Université de Montréal

Gestion des ressources dans les réseaux cellulaires sans fil

Par
Apollinaire Nadembéga

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ doctor (Ph.D.)
en Informatique

Décembre, 2013

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée :

Gestion des ressources dans les réseaux cellulaires sans fil

Présentée par :

Apollinaire Nadembéga

a été évaluée par un jury composé des personnes suivantes :

El Mostapha Aboulhamid, président-rapporteur

Abdelhakim Hafid, directeur de recherche

Samuel Pierre, membre du jury

Roch Glitho, examinateur externe

……., représentant du doyen

Résumé

L'émergence de nouvelles applications et de nouveaux services (tels que les applications multimédias, la voix-sur-IP, la télévision-sur-IP, la vidéo-sur-demande, etc.) et le besoin croissant de mobilité des utilisateurs entrainent une demande de bande passante de plus en plus croissante et une difficulté dans sa gestion dans les réseaux cellulaires sans fil (WCNs), causant une dégradation de la qualité de service. Ainsi, dans cette thèse, nous nous intéressons à la gestion des ressources, plus précisément à la bande passante, dans les WCNs.

Dans une première partie de la thèse, nous nous concentrons sur la prédiction de la mobilité des utilisateurs des WCNs. Dans ce contexte, nous proposons un modèle de prédiction de la mobilité, relativement précis qui permet de prédire la destination finale ou intermédiaire et, par la suite, les chemins des utilisateurs mobiles vers leur destination prédite. Ce modèle se base sur : (a) les habitudes de l'utilisateur en terme de déplacements (filtrées selon le type de jour et le moment de la journée) ; (b) le déplacement courant de l'utilisateur ; (c) la connaissance de l'utilisateur ; (d) la direction vers une destination estimée ; et (e) la structure spatiale de la zone de déplacement. Les résultats de simulation montrent que ce modèle donne une précision largement meilleure aux approches existantes.

Dans la deuxième partie de cette thèse, nous nous intéressons au contrôle d'admission et à la gestion de la bande passante dans les WCNs. En effet, nous proposons une approche de gestion de la bande passante comprenant : (1) une approche d'estimation du temps de transfert intercellulaire prenant en compte la densité de la zone de déplacement en terme d'utilisateurs, les caractéristiques de mobilité des utilisateurs et les feux tricolores ; (2) une approche d'estimation de la bande passante disponible à l'avance dans les cellules prenant en compte les exigences en bande passante et la durée de vie des sessions en cours ; et (3) une approche de réservation passive de bande passante dans les cellules qui seront visitées pour les sessions en cours et de contrôle d'admission des demandes de nouvelles sessions prenant en compte la mobilité des utilisateurs et le comportement des cellules. Les résultats de simulation indiquent que cette approche réduit largement les ruptures abruptes de sessions en cours, offre un taux de refus de nouvelles demandes de connexion acceptable et un taux élevé d'utilisation de la bande passante.

Dans la troisième partie de la thèse, nous nous penchons sur la principale limite de la première et deuxième parties de la thèse, à savoir l'évolutivité (selon le nombre d'utilisateurs) et proposons une plateforme qui intègre des modèles de prédiction de mobilité avec des modèles de prédiction de la bande passante disponible. En effet, dans les deux parties précédentes de la thèse, les prédictions de la mobilité sont effectuées pour chaque utilisateur. Ainsi, pour rendre notre proposition de plateforme évolutive, nous proposons des modèles de prédiction de mobilité par groupe d'utilisateurs en nous basant sur : (a) les profils des utilisateurs (c'est-à-dire leur préférence en termes de caractéristiques de route) ; (b) l'état du trafic routier et le comportement des utilisateurs ; et (c) la structure spatiale de la zone de déplacement. Les résultats de simulation montrent que la plateforme proposée améliore la performance du réseau comparée aux plateformes existantes qui proposent des modèles de prédiction de la mobilité par groupe d'utilisateurs pour la réservation de bande passante.

Mots-clés : Réseaux cellulaires sans fil ; qualité de service ; prédiction de la mobilité ; réservation de la bande passante ; priorisation du transfert intercellulaire ; contrôle d'admission.

Abstract

The emergence of new applications and services (e.g., multimedia applications, voice over IP and IPTV) and the growing need for mobility of users cause more and more growth of bandwidth demand and a difficulty of its management in Wireless Cellular Networks (WCNs). In this thesis, we are interested in resources management, specifically the bandwidth, in WCNs.

In the first part of the thesis, we study the user mobility prediction that is one of key to guarantee efficient management of available bandwidth. In this context, we propose a relatively accurate mobility prediction model that allows predicting final or intermediate destinations and subsequently mobility paths of mobile users to reach these predicted destinations. This model takes into account (a) user's habits in terms of movements (filtered according to the type of day and the time of the day); (b) user's current movement; (c) user's contextual knowledge; (d) direction from current location to estimated destination; and (e) spatial conceptual maps. Simulation results show that the proposed model provides good accuracy compared to existing models in the literature.

In the second part of the thesis, we focus on call admission control and bandwidth management in WCNs. Indeed, we propose an efficient bandwidth utilization scheme that consists of three schemes: (1) handoff time estimation scheme that considers navigation zone density in term of users, users' mobility characteristics and traffic light scheduling; (2) available bandwidth estimation scheme that estimates bandwidth available in the cells that considers required bandwidth and lifetime of ongoing sessions; and (3) passive bandwidth reservation scheme that passively reserves bandwidth in cells expected to be visited by ongoing sessions and call admission control scheme for new call requests that considers the behavior of an individual user and the behavior of cells. Simulation results show that the proposed scheme reduces considerably the handoff call dropping rate while maintaining acceptable new call blocking rate and provides high bandwidth utilization rate.

In the third part of the thesis, we focus on the main limitation of the first and second part of the thesis which is the scalability (with the number of users) and propose a framework,

together with schemes, that integrates mobility prediction models with bandwidth availability prediction models. Indeed, in the two first contributions of the thesis, mobility prediction schemes process individual user requests. Thus, to make the proposed framework scalable, we propose group-based mobility prediction schemes that predict mobility for a group of users (not only for a single user) based on users' profiles (i.e., their preference in terms of road characteristics), state of road traffic and users behaviors on roads and spatial conceptual maps. Simulation results show that the proposed framework improves the network performance compared to existing schemes which propose aggregate mobility prediction bandwidth reservation models.

Keywords: Wireless cellular networks; Quality of service (QoS); Mobility prediction; Bandwidth reservation; Prioritized handoff; Call admission control.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

3GPP	3rd Generation Partnership Project
ABE	Available Bandwidth Estimation scheme
AHTES	Aggregate Handoff Times Estimation Scheme
APPM	Aggregate Path Prediction Model
BTS/BS	Base Transceiver Station / Base Station
CAC	Call Admission Control
CDF	Cumulative Distribution Function
CTL	Controller
DAMP	Destination And Mobility path Prediction
DPM	Destination Prediction Model
DSZ	Dense Sub-Zone
ECaC	Efficient Call admission Control scheme
FVL	Frequently Visited Location
GPS	Global Positioning System
HTE	Handoff Times Estimation scheme
HTEMOD	Handoff Time Estimation MODel
IP	Internet Protocol
IPMBRF	Integrated Predictive Mobile-oriented Bandwidth Reservation Framework
LTE	Long Term Evolution
MN	Mobile Networks
MPBR	Mobility Prediction aware Bandwidth-Reservation scheme
NM	Navigation Map
NS	Network System
PDF	Probability Distribution Function
PPM	Path Prediction Model
QdS	Qualité de Service
QoE	Quality of Experience
QoS	Quality of Service
SQL	Structured Query Language

UC	User Contextual
UE	User Equipment
WCN	Wireless Cellular Network
WiMAX	Worldwide Interoperability for Microwave Access

Glossaire des traductions

Backbone	Épine dorsale du réseau
Bandwidth	Bande passante
Base Station	Station de base
Call Admission Control	Contrôle d'admission des sessions
Cumulative Distribution Function	Fonction cumulative de la distribution de probabilité
Dense Sub-Zone	Sous-zone très dense
Frequently Visited Location	Localité fréquemment visitée par l'utilisateur
Global Positioning System	Système de localisation globale
Handoff	Transfert intercellulaire
Internet Protocol	Protocole d'Internet
Mobile Networks	Réseaux mobiles
Navigation Map	Carte de navigation
Probability Distribution Function	Fonction de la distribution de probabilité
Quality of service	Qualité de service
User Contextual	Information contextuelle sur l'utilisateur
Wireless Cellular Network	Réseau cellulaire sans fil

À ma mère, Maria Lucie Nadembéga, née Sawadogo

Remerciements

Je remercie mon directeur de recherche Abdelhakim Hafid, professeur à l'Université de Montréal, pour la qualité de son encadrement et l'intérêt qu'il a démontré à l'égard de cette thèse. Je le remercie de même pour ses encouragements, conseils et directives pertinentes qui ont été d'une grande importance pour la réalisation de ce projet de recherche. Enfin, je le remercie surtout pour sa patience, sa compréhension et son aide face à mes difficultés de rédaction d'articles.

Mes remerciements s'adressent également au Dr. Tarik Taleb, chercheur à NEC Europe, pour les discussions et critiques fructueuses et les précieux conseils dont il nous a gratifiés au cours de cette thèse.

Je remercie ma conjointe, ma famille et mes amis qui m'ont donné la force, l'ambition et le soutien nécessaires pour arriver au bout de ce projet.

Je tiens aussi à remercier les collègues du Laboratoire des Réseaux de Communications (LRC) de l'Université de Montréal, en particulier Dr. Mustapha Boushaba, pour la collaboration et le soutien que j'ai reçus.

Enfin, je remercie toutes celles et tous ceux qui ont participé de près ou de loin à l'accomplissement de cette thèse.

Chapitre 1 : Introduction

Dans ce chapitre, nous présentons le contexte général et les motivations de la thèse. Ensuite, nous décrivons les problèmes de gestion des ressources dans les réseaux cellulaires sans fil ainsi que les contributions de cette thèse. Enfin, nous donnons un aperçu de l'organisation de la thèse.

1.1. Contexte général

De nos jours, les réseaux cellulaires sans fil (WCNs) connaissent un essor considérable [1]. Cela est dû à l'étendue de la couverture, la possibilité d'accéder à Internet n'importe où et n'importe quand, la facilité d'utilisation (téléphone mobile et abonnement auprès d'un opérateur) et la capacité de transmission due à l'évolution de la technologie radio et de la communication sans fil. Par exemple, le Long Term Evolution-Advanced [2, 3] offre des débits descendants pouvant atteindre 299.552 Mbits/s et des débits montants de 75.376 Mbits/s avec un délai de transmission de moins de 10 ms [2]. Les WCNs assurent une connectivité mobile transparente aux utilisateurs ; c'est donc un réseau mobile. A cet effet, il est nécessaire de faire la distinction entre un réseau sans fil et un réseau mobile. Un réseau sans fil [2, 4-7] peut se définir comme un réseau dans lequel les nœuds (utilisateurs, routeurs, points d'accès, etc.) ne sont pas reliés entre eux par des médiums de communication filaires. Un réseau mobile est un réseau sans fil, mais segmenté ou subdivisé dans lequel les utilisateurs (appelés utilisateurs mobiles) sont mobiles et s'attachent, de façon arbitraire, à des points d'accès dans l'infrastructure de routage [8] ; tel est le cas des réseaux maillés sans fil [5, 6], des réseaux véhiculaires [7] et des WCNs [2]. L'avènement des WCNS a engendré la téléphonie mobile ou cellulaire. La téléphonie mobile ou cellulaire est un moyen de télécommunications par téléphone sans fil (téléphone mobile) qui s'est répandu dans les années 1990.

Généralement, l'architecture des WCNs est composée de deux parties : la partie accès radio du réseau et la partie épine dorsale ou cœur du réseau. La figure 1 montre une image de l'architecture simplifiée d'un WCN. Dans le cas du LTE [3], la partie épine dorsale est responsable de tous les contrôles reliés aux utilisateurs (authentification, mobilité, identification des classes de QdS, établissement de la connexion), de la gestion de la

tarification, de la gestion des abonnés (utilisateurs qui ont souscrit un abonnement chez un opérateur de téléphonie mobile), de la gestion des adresses IP et du transfert des paquets vers les destinataires [3].

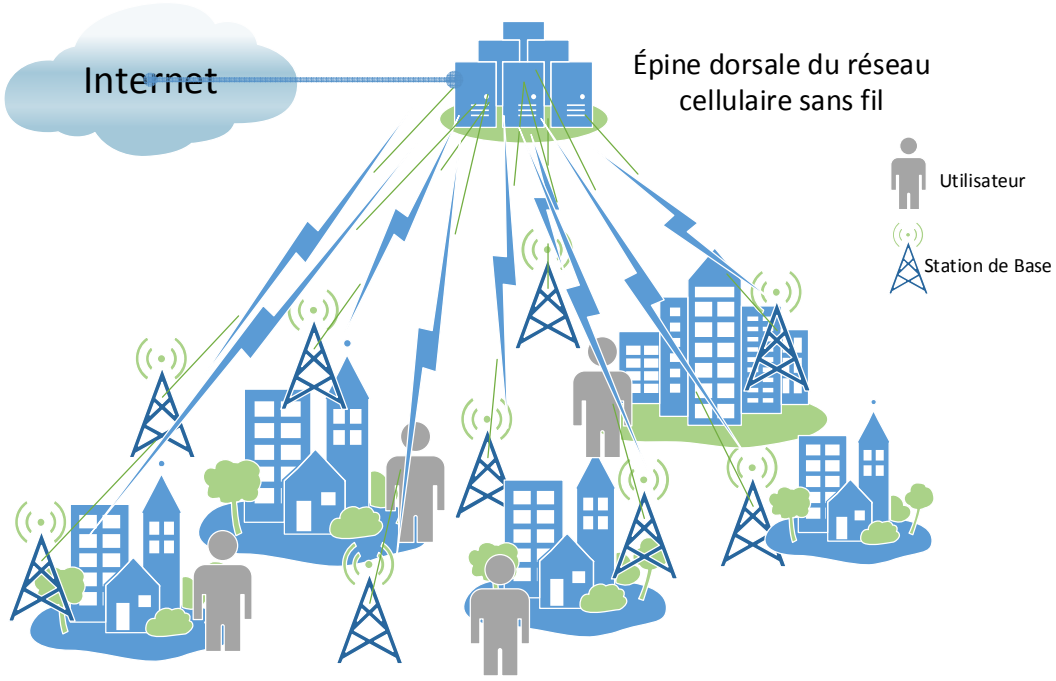


Figure 1 : Réseau cellulaire sans fil

La partie accès radio utilise les ondes radioélectriques (dites ondes radio ou hertziennes) qui sont des ondes électromagnétiques dont la fréquence est par convention inférieure à 300 GHz. Elles se propagent dans l'espace de façon naturelle. Dans le cas du LTE [3], la partie accès radio est responsable de tous les contrôles reliés aux ondes radios (admission, mobilité, planification et allocation des ressources aux utilisateurs), de la compression de l'en-tête IP, du cryptage des données transmises via l'accès radio et de la connectivité avec la partie épine dorsale du réseau [3]. Il joue le rôle d'intermédiaire entre les utilisateurs et l'épine dorsale. La partie accès radio est composée de points d'accès appelés « station de base ». Les stations de base permettent aux utilisateurs d'être reliés à la partie accès radio du réseau. Dans le cas du LTE, les stations de base sont connectées entre elles (possibilité de communication directe entre stations de base sans passer par l'épine dorsale) et aussi avec la partie épine dorsale. Chaque station de base est accessible par un utilisateur à partir d'une position se trouvant dans

sa couverture radio. La couverture radio engendrée par une station de base est appelée cellule. Les cellules sont des zones circulaires se chevauchant afin de couvrir une zone géographique. Chaque cellule possède une capacité en bande passante limitée ; cependant, plus le rayon d'une cellule est petit, plus sa capacité en bande passante est élevée [3]. C'est la raison pour laquelle, dans les zones urbaines, la taille des cellules avoisine quelques centaines de mètres, tandis que de vastes cellules d'une trentaine de kilomètres permettent de couvrir les zones rurales. Chaque cellule est entourée de six (6) cellules voisines, d'où la représentation d'une cellule par un hexagone. La figure 2 indique une illustration d'une cellule de WCN.

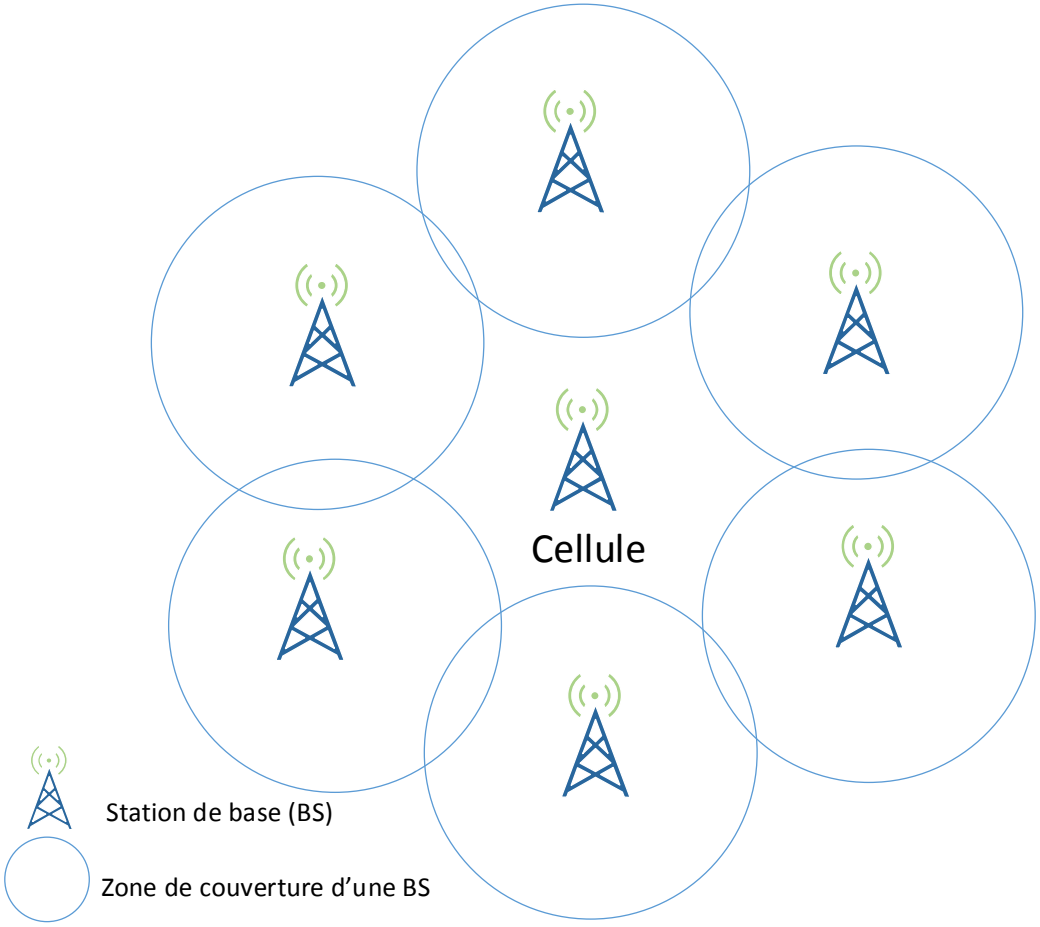


Figure 2 : Illustration d'une cellule de réseau cellulaire sans fil

Une amélioration importante dans les WCNs (cas du LTE [2, 3]) est la possibilité d'utiliser les mêmes fréquences hertziennes/radios dans toutes les cellules ; ce qui n'était pas possible avec

le réseau mobile de seconde génération. En conséquence, cela permet d'avoir une bande passante plus importante et plus de débit dans chaque cellule.

L'évolution des WCNs a conduit à une nouvelle ère de services et d'applications mobiles ainsi qu'à des terminaux mobiles (p.ex., les téléphones cellulaires intelligents et les tablettes numériques) équipés de processeurs plus puissants et de capacité de stockage plus grande. Cette évolution fait du WCN un réseau attrayant pour les utilisateurs. Par conséquent, cela cause une augmentation rapide du nombre d'utilisateurs des WCNs, laquelle augmentation engendre un fort taux d'utilisation de ces ressources radios entrainant son insuffisance. Cela suscite un grand intérêt auprès de la communauté de la recherche dans le but de proposer des solutions pour améliorer ses performances. Effectivement, vu que la capacité des cellules en bande passante est limitée, la demande due au grand nombre d'utilisateurs cause une dégradation de la qualité de service ; donc une dégradation de la performance des WCNs. Cette dégradation de performance est accentuée par la mobilité des utilisateurs.

C'est dans ce contexte des WCNs que se situe ce travail de thèse dont l'objectif est de contribuer à mettre en place des solutions pour gérer l'utilisation de la bande passante disponible dans les WCNs. Ces solutions doivent réduire le taux de rupture inopinée de session en cours tout en maintenant un taux acceptable de refus de demandes de nouvelles sessions. Une session est une période de communication établie entre deux entités (p.ex., utilisateurs, applications, services) qui débute par la connexion et qui se termine par la déconnexion.

1.2. Motivations

À l'origine limité au domaine de la recherche, Internet s'est ouvert, de nos jours, à des domaines multiples faisant de lui le moyen de communication distant le plus utilisé. Cette ouverture à des domaines diversifiés a nécessité plusieurs innovations technologiques, aussi bien au niveau matériel qu'au niveau logiciel, afin de répondre aux besoins grandissant des internautes (utilisateurs d'Internet). Grâce à l'évolution des WCNs et l'avènement des terminaux mobiles intelligents, l'accès à Internet via les WCNs est devenu de plus en plus incontournable dans le quotidien des êtres humains. La possibilité pour un utilisateur de

demeurer connecté à Internet tout en étant en déplacement a contribué à accroitre très rapidement le nombre d'internautes via des WCNs. Aussi, avec (1) l'émergence de nouvelles applications et de nouveaux services tels que les applications multimédias, la voix-sur-IP et la vidéo-sur-demande ; et (2) la capacité des nouvelles technologies dans les WCNs telles que le Worldwide Interoperability for Microwave Access (WiMAX) et le LTE de fournir des capacités de transmission répondant aux besoins de ces applications et services [9-14], le nombre des internautes via les WCNs ne fait qu'augmenter : plus de 2.7 milliards d'internautes et 6.8 milliards d'abonnements mobiles en 2013 dans le monde selon l'Union Internationale des Télécommunications [15]. Cette augmentation des utilisateurs mobiles (de 20% à 96.2% entre 2003 et 2013 [15]) induit une demande de bande passante de plus en plus croissante dans les cellules.

Malheureusement, cette bande passante est limitée et son insuffisance est la raison principale de la dégradation de la qualité de service (QdS) fournie aux utilisateurs. La QdS est la capacité d'un réseau à véhiculer dans de bonnes conditions le flux de données/paquets afin de permettre une bonne qualité de communication aux utilisateurs. La QdS est caractérisée par un certain nombre de paramètres : la disponibilité, le débit, les délais de transmission, le gigue (différence de délai de transmission de bout en bout entre des paquets choisis dans un même flux de paquets), le taux de perte de paquets, le taux de rupture inopinée de session et le taux de refus de nouvelle session. La QdS a pour but d'optimiser les ressources du réseau et d'assurer de bonnes performances aux applications et services. En effet, le manque de bande passante dans une cellule du réseau provoque une rupture inopinée des sessions, ce qui conduit à l'arrêt forcé et immédiat des applications et services en cours ou un refus des demandes de nouvelles sessions. Cette dégradation de la QdS est renforcée par le caractère mobile des utilisateurs. Effectivement, la rupture inopinée de la session se produit très souvent pendant les transferts intercellulaires des utilisateurs en déplacement (utilisateurs mobiles). Un transfert intercellulaire est le changement de point d'accès/attachement/cellule qu'un utilisateur mobile effectue fréquemment pendant son déplacement [16]. Notons que lorsqu'un utilisateur mobile en communication (p. ex., connecté, via un WCN, à des applications ou des services hébergés sur des serveurs du réseau Internet) se déplace, il traverse des cellules qui disposent de quantités en bande passante différentes. Alors, lorsque cet utilisateur arrive dans une nouvelle

cellule où la quantité de bande passante disponible n'est pas suffisante pour le bon fonctionnement de ses applications et services en cours d'exécution, le WCN est dans l'obligation de mettre fin à toutes ou certaines de ces sessions. Ce qui provoque immédiatement la fin des applications et services. Il peut arriver également que le WCN réduise la quantité de bande passante qui était allouée à l'utilisateur dans la cellule précédente due à une insuffisance de la quantité de bande passante disponible dans la nouvelle cellule. Ce qui provoque une réduction de la QdS même si les sessions (applications et services) ne sont pas arrêtées. Par conséquent, il se pose donc un problème de gestion efficace de la bande passante disponible dans les cellules des WCNs due principalement à la mobilité des utilisateurs. Le caractère mobile imprévisible des utilisateurs ne permet pas aux WCNs de mieux s'organiser pour accroitre le niveau de satisfaction des utilisateurs. Les conséquences de ce manque de gestion efficace de la bande passante disponible sont (1) la non satisfaction de la QdS exigée par utilisateurs (p.ex., la rupture inopinée et abrupte/forcée de la session mettant fin aux applications et aux services en cours exécution sans préavis) ; et (2) la perte de revenus pour les opérateurs/fournisseurs de WCNs (p.ex., la sous-utilisation de la bande passante disponible et les utilisateurs insatisfaits qui vont changer de fournisseurs).

Alors, le support de la QdS fournie aux utilisateurs dans les WCNs passe par une bonne gestion de la bande passante disponible et un contrôle efficace de son allocation. Cette gestion demeure un des défis majeurs dans ce type de réseau [17-19]. Puisqu'elle est étroitement liée à la mobilité des utilisateurs, il faudrait une prise en compte de celle-ci pour une gestion efficace. Donc, prédire les déplacements des utilisateurs avec une précision acceptable est la clé pour supporter la QdS [20]. Effectivement, connaitre à l'avance les chemins qui seront empruntés et les temps de transferts intercellulaires des utilisateurs mobiles permettra de savoir à l'avance les quantités de bande passante qui seront disponibles dans les cellules [17] dans un futur proche. Cela conduira à mieux contrôler l'allocation de la bande passante aux nouvelles demandes de sessions et de réserver des quantités raisonnables de bande passante pour les utilisateurs qui arriveront plus tard dans les cellules. Cependant, le processus de réservation devra prendre en compte les temps d'arrivée des utilisateurs dans les cellules afin d'éviter une sous-utilisation et une réservation abusive de la bande passante disponible.

1.3. Description des problèmes

À la lumière des défis cités dans la précédente section, la première problématique de la thèse est la prédiction des déplacements des utilisateurs ; c'est-à-dire, déterminer avec une précision acceptable les différents chemins qui seront empruntés par les utilisateurs mobiles. Savoir à l'avance les chemins des utilisateurs est actuellement irréalisable, mais elle peut être déduite avec une certaine probabilité. Plusieurs études [20-25] ont été faites dans ce sens, mais elles sont jugées très peu réalistes car fondées sur de nombreuses suppositions et occultant la prise en compte de certains critères importants.

La seconde problématique de la thèse est celle de l'estimation des temps de transferts intercellulaires des utilisateurs durant leurs déplacements et la gestion de l'allocation de la bande passante. La gestion de l'allocation de la bande passante se base sur les résultats de l'estimation des temps de transfert intercellulaire pour estimer, à l'avance les quantités de bande passante disponibles dans les cellules. L'estimation des temps de transferts intercellulaires est la problématique la plus complexe à résoudre. Elle nécessite une prise en compte de plusieurs contraintes liées aux règles et aux signalisations de la circulation routière, aux lois de la physique sur le déplacement d'un corps, à la topologie des routes et aux incidents sur les routes pour avoir une meilleure précision. Actuellement, très peu d'études sont menées sur le sujet à cause de sa complexité. En outre, les études [26-28] qui ont été faites sur ce sujet considèrent la vitesse ou l'arrêt à un panneau d'arrêt ou le temps d'arrêt à un carrefour comme des éléments ou événements aléatoires. Ce qui rend les résultats de ces études moins réalistes.

La troisième problématique de la thèse est celle de l'intégration de la prédiction de la mobilité (chemins et temps de transferts intercellulaires) avec la gestion de l'allocation de la bande passante disponible. Elle concerne également une architecture d'ensemble montrant l'interaction entre les entités en vue de supporter la QdS. Spécifiquement, la troisième problématique de la thèse s'intéresse à l'évolutivité des modèles de mobilité avec le nombre d'utilisateurs. Une prédiction individuelle de la mobilité, dans le cas d'un grand nombre d'utilisateurs, causerait un surplus de charge en termes de traitement et engendrerait des délais d'attente de connexion plus élevés que de coutume. En un mot, nous nous intéressons dans la troisième problématique à supporter l'utilisation efficace de la bande passante disponible en

nous basant sur la prédiction globale de la mobilité des utilisateurs. L'objectif visé dans la résolution de la troisième problématique est de réserver efficacement la bande passante pour les sessions (applications et services) en cours d'exécution afin de réduire le taux de leur rupture inopinée. C'est aussi de mieux planifier l'acceptation des nouvelles demandes de sessions afin de maintenir un taux acceptable de leur refus et finalement d'éviter une sous-utilisation de la bande passante disponible. Notons que l'étude effectuée dans cette thèse est caractérisée par son aspect de satisfaction pour les utilisateurs avec une réduction des ruptures inopinées de la connectivité des sessions et un taux acceptable de demandes de nouvelles sessions refusées. Elle vise aussi un aspect économique pour les opérateurs des WCNs et les fournisseurs d'applications et de services mobiles qui verront accroitre leur clientèle avec l'augmentation du niveau de satisfaction des utilisateurs ; ce qui permettra d'accroitre leur chiffre d'affaires.

1.4. Contributions de la thèse

Dans cette thèse, nous nous penchons sur les trois problématiques de la QdS dans les WCNs mentionnées dans la section précédente. Ces trois problématiques sont : (1) la prédiction des chemins des utilisateurs ; (2) l'estimation des temps de transferts intercellulaires et leur prise en compte dans la gestion de l'allocation de la bande passante disponible ; et (3) l'intégration de la prédiction globale de la mobilité (chemins et temps de transferts intercellulaires) avec la gestion de l'allocation de la bande passante dans une architecture. Nous proposons trois solutions dont chacune fait l'objet d'une contribution.

Pour la première contribution, nous proposons un modèle de prédiction de mobilité que nous appelons *Destination And Mobility path Prediction* (DAMP). DAMP est composé de deux modèles : (1) *Destination Prediction Model* (DPM) et (2) *Path Prediction Model* (PPM). Plus spécifiquement, DPM permet de déduire la destination de l'utilisateur mobile en prenant en compte (a) les habitudes de l'utilisateur en termes de fréquence de visites de certains lieux ; (b) la direction du mouvement de l'utilisateur ; et (c) la connaissance de l'utilisateur. En effet, faisant usage de la direction du mouvement, DPM détermine les lieux potentiels pouvant être la destination de l'utilisateur et regroupe ceux pouvant être atteints en utilisant un même chemin pendant un temps prédéfini ; DPM considère donc un groupe de destinations comme

étant la destination de notre utilisateur mobile : il réduit de ce fait le champ de recherche de la destination et les risques d'erreur que pourrait entrainer l'usage des données historiques et de la connaissance de l'utilisateur ; il augmente aussi la précision de la prédiction du chemin. DPM évalue la probabilité qu'un groupe de lieux soit la destination de l'utilisateur en utilisant les données historiques sur les fréquences de visite de ces lieux par l'utilisateur (filtrées selon le type de jour et le moment de la journée pour prendre en compte les habitudes de l'utilisateur) et le second ordre de la chaîne de Markov. Ensuite, DPM utilise les travaux de N. Samaan et A. Karmouch [29] pour évaluer l'évidence qu'un groupe de lieux soit la destination de l'utilisateur en s'appuyant sur la connaissance de ce dernier. Enfin, DPM somme les valeurs obtenues après les deux évaluations en associant un poids croissant à la probabilité selon la quantité de données historiques ; DPM choisit le groupe de lieux ayant la plus grande valeur de la somme. En se basant sur la destination estimée, PPM (second modèle de DAMP) estime le chemin qui sera emprunté par l'utilisateur pour atteindre la destination estimée (groupe de lieux) par DPM. PPM prend en compte (a) les habitudes de l'utilisateur en termes de fréquence de rue utilisée pour atteindre une destination donnée ; (b) la direction vers la destination estimée ; et (c) la structure spatiale de la zone de déplacement qui est formée d'intersections de routes et de segments de routes ; nous définissons une intersection de routes (carrefour) comme étant un lieu de jonction ou de croisement de deux ou plusieurs routes et un segment de route comme une portion de route reliant deux intersections de routes adjacentes (juxtaposées, contigües) ; un segment de route A est dit adjacent à un autre segment de route B s'il est possible d'accéder immédiatement à A après avoir transité par B. Plus précisément, à chaque carrefour (intersection de routes) PPM détermine le prochain segment de route qui sera emprunté par l'utilisateur sachant la source de son déplacement et sa destination prévue ; PPM choisit celui qui a la plus grande probabilité calculée à partir des données historiques de mobilité filtrées selon le type de jour et le moment de la journée. Le choix de ce segment de route est fait parmi un groupe de segments de route adjacents au segment de route précédemment choisi ; ce groupe est formé en s'appuyant sur la direction vers la destination. PPM répète cette opération jusqu'à ce que la destination soit atteinte. La séquence de segments de route obtenue constitue le chemin estimé vers la destination estimée. La deuxième contribution est consacrée aux problèmes d'estimation des temps de transferts intercellulaires et de la gestion d'allocation de la bande passante disponible. Pour ce faire,

nous proposons une approche de gestion efficace de la bande passante disponible, appelée *Mobility Prediction-aware Bandwidth Reservation scheme* (MPBR). MPBR est constitué de trois systèmes : (1) *Handoff Time Estimation scheme* (HTE) (2) *Available Bandwidth Estimation scheme* (ABE) et (3) *Efficient Call admission Control scheme* (ECaC). HTE estime un intervalle de temps pendant lequel un utilisateur pourrait effectuer un transfert intercellulaire durant son déplacement. HTE suppose que le chemin de l'utilisateur vers sa destination est déjà connu grâce aux travaux de la contribution précédente [21, 30]. HTE, qui est une extension de notre proposition HTEMOD [31], prend en considération : (a) la densité de la zone de déplacement ; (b) le comportement du déplacement de l'utilisateur sur la route (c'est-à-dire l'accélération, vitesse maximale et la décélération) ; et (c) la gestion des feux tricolores. Plus spécifiquement, HTE définit la fonction de la distribution de probabilité (PDF), qui diffère selon la densité, puis la fonction cumulative de la distribution de probabilité (CDF) du temps mis par un utilisateur pour atteindre un point du transfert intercellulaire. Puis, choisissant un niveau de probabilité et l'inverse du CDF défini, HTE estime l'intervalle de temps où celui-ci est susceptible d'effectuer son transfert intercellulaire. Une fois les intervalles de temps de transferts intercellulaires des utilisateurs obtenus, MPBR utilise ABE dont le but est d'estimer, à l'avance (p. ex., 30 minutes), la quantité de bande passante qui pourrait être disponible dans chaque cellule ; ABE prend en compte (a) les sessions (c'est-à-dire applications et services) entrantes et sortantes des cellules ; (b) la durée restante des sessions en cours ; et (c) les intervalles des temps de transferts intercellulaires (supposés connus grâce à HTE). Enfin, sachant à l'avance la quantité estimée de bande passante disponible dans les cellules, MPBR fait recours à ECaC pour contrôler l'admission des demandes de sessions. ECaC prend en compte, en plus de la mobilité des utilisateurs, (a) le comportement individuel des cellules ; (b) la totalité du chemin de la position courante vers la destination des utilisateurs ; et (c) la priorité des sessions en cours sur les nouvelles demandes de sessions.

La troisième contribution est consacrée au problème de l'intégration de la prédiction globale de la mobilité des utilisateurs avec la gestion de l'allocation de la bande passante, le tout dans une architecture indiquant les interactions entre les entités impliquées par l'intégration. Cette contribution répond à la principale limite des deux précédentes

contributions où la prédiction de la mobilité est effectuée individuellement pour chaque utilisateur. Nous proposons une plateforme d'intégration, appelée *Integrated Predictive Mobile-oriented Bandwidth Reservation Framework (IPMBRF).* Les modèles globaux de prédiction de la mobilité permettent de déduire les chemins que pourraient emprunter un groupe d'utilisateurs et les temps de transferts intercellulaires de ceux-ci en un seul processus. Faisant usage des modèles globaux de prédiction de la mobilité, le modèle de prédiction de la bande passante disponible aide à mieux accommoder les sessions des utilisateurs pendant leurs déplacements. Premièrement, nous proposons un modèle de prédiction global de chemin (pour un groupe d'utilisateurs) entre des sous-zones basé sur le profil des utilisateurs. Celui-ci considère (a) les préférences des utilisateurs en termes de caractéristiques de route (par exemple autoroute, route à sens unique, route à plusieurs voies, route sans panneau d'arrêt, et route sans feu tricolore) ; (b) la structure spatiale de la zone de navigation ; et (c) les destinations prévues qui seront fournies par le DPM de la première contribution. Le modèle d'estimation global du temps de transfert intercellulaire (pour un groupe d'utilisateurs) construit une table de temps de transit pour chaque segment de route en prenant en compte (a) l'état du trafic routier et (b) le comportement en termes de circulation routière des utilisateurs. Sur la base des tables de temps de transit des segments de route, les temps de transferts intercellulaires des utilisateurs sont calculés. L'architecture de la plateforme proposée se compose de l'équipement de l'utilisateur et d'un contrôleur. L'équipement de l'utilisateur est en charge de la détermination de la destination et du chemin prévu en cas de densité faible dans la zone de navigation. Le contrôleur, lui, est en charge de la détermination des chemins en cas de densité élevée, de l'estimation des temps de transferts intercellulaires, de la prédiction des quantités de bande passante disponibles dans les cellules et du contrôle d'admission des sessions.

1.5. Organisation de la thèse

La thèse est organisée comme suit. Après ce chapitre introductif, le chapitre 2 fait une revue de la littérature sur la prédiction de la mobilité et le provisionnement de la qualité de service dans les WCNs. Ensuite, le chapitre 3 présente la première contribution sur la prédiction de la mobilité des utilisateurs dans les WCNs. Le chapitre 4 élabore sur la deuxième

contribution qui est l'estimation des temps de transferts intercellulaires et de la gestion de l'allocation de la bande passante afin de supporter la qualité de service dans les WCNs. Dans le chapitre 5, nous nous penchons sur la troisième contribution. Celle-ci est la plateforme d'intégration des modèles de prédiction de la mobilité (chemin et temps de transferts intercellulaires) pour un groupe d'utilisateurs en une seule requête avec le modèle d'estimation de la bande passante disponible. Finalement, le chapitre 6 trace les conclusions de cette thèse et identifie quelques pistes de recherche pour les travaux futurs.

La Figure 3 illustre un organigramme des contributions et de l'organisation de cette thèse.

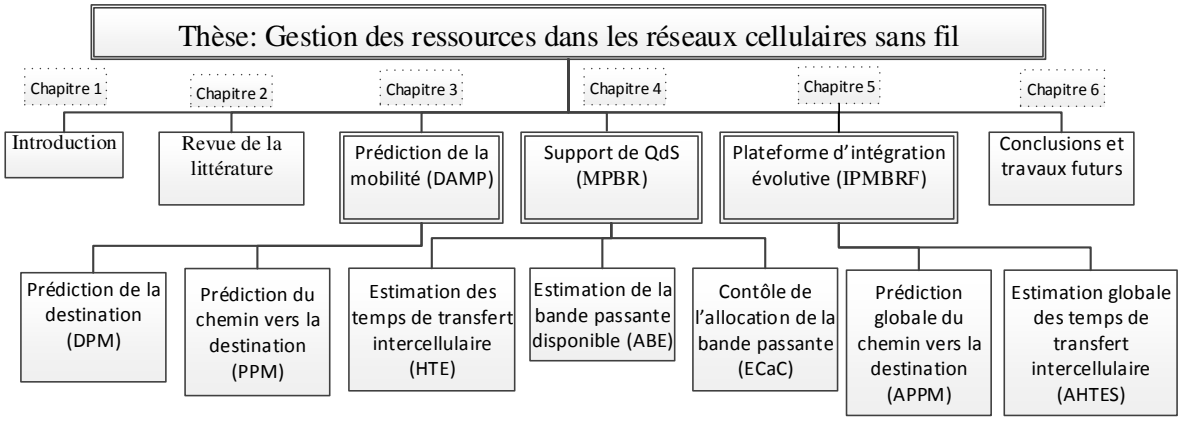


Figure 3 : Organigramme de l'organisation de la thèse.

1.6. Publications de la thèse

La liste des articles de revues et des articles de conférences rédigés au cours de cette thèse est la suivante :

1. A. Nadembéga, A. Hafid and T. Taleb. *An Integrated Predictive Mobile-Oriented Bandwidth-Reservation Framework to Support Mobile Multimedia*, soumis à IEEE TWC, 2013.
2. A. Nadembéga, A. Hafid and T. Taleb. *Mobility Prediction-aware Bandwidth Reservation Scheme for Mobile Networks,* soumis à IEEE TVT, 2013.
3. A. Nadembéga, A. Hafid and T. Taleb. *DAMP: A Destination & Mobility Path Prediction Scheme,* soumis à IEEE TVT, 2013.

4. A. Nadembéga, A. Hafid and T. Taleb. *A Framework for Mobility Prediction and High Bandwidth Utilization to Support Mobile Multimedia Streaming,* IEEE ANTS, Chennai, India, Dec. 2013.

5. A. Nadembéga, A. Hafid and T. Taleb. *Handoff Time Estimation Model for Vehicular Communications*, IEEE ICC, Budapest, Hungary, Jun. 2013.

6. A. Nadembéga, A. Hafid and T. Taleb. *A Destination Prediction Model based on Historical Data, Contextual Knowledge and Spatial Conceptual Maps,* IEEE ICC, Ottawa, Ontario, Canada, Jun. 2012.

7. A. Nadembéga, A. Hafid and T. Taleb. *A Path Prediction Model to Support Mobile Multimedia Streaming*, IEEE ICC, Ottawa, Ontario, Canada, Jun. 2012.

8. T. Taleb, A. Hafid, and A. Nadembéga. *Mobility-Aware Streaming Rate Recommendation System*, IEEE Globecom, Houston, Texas, USA, Nov. 2011.

Chapitre 2 : Revue de la littérature

Dans ce chapitre, nous présentons un aperçu sur la mobilité dans les WCNs et la gestion des ressources. Dans ce cadre, nous relevons les travaux réalisés sur la prédiction de la mobilité des utilisateurs dans les WCNs et l'estimation de leurs temps de transfert intercellulaires. Par la suite, nous exposons les travaux sur le contrôle d'admission des sessions/appels (applications et services mobiles) et la gestion de la bande passante.

2.1. La mobilité dans les WCNs

Le concept de mobilité ne dépend pas d'un type spécifique de réseaux sans fil ; elle est perçue comme l'ensemble des aspects qui sont liés aux déplacements des utilisateurs ou des routeurs dans le réseau. Dans notre présentation de la mobilité, nous prendrons en compte la mobilité des utilisateurs en tant que nœuds terminaux et non la mobilité des équipements (p.ex. les routeurs) formant l'architecture du réseau Considérant la mobilité des utilisateurs, le processus le plus important de la mobilité dans les WCNs est le transfert intercellulaire. Le processus du transfert intercellulaire consiste à ce qu'un terminal mobile maintienne sa session en cours, lors d'un déplacement qui l'amène à changer de cellule. En effet, lorsque le signal de transmission entre un terminal mobile et une station de base s'affaiblit, le terminal mobile cherche une autre station de base disponible dans une autre cellule qui est capable d'assurer à nouveau la communication dans les meilleures conditions. La figure 4 montre une illustration simplifiée du transfert intercellulaire. De façon générale, des notes techniques et organisationnelles concernant la mobilité via Internet peuvent être consultées dans [8, 32-43].

Au regard de la relation d'appartenance entre un utilisateur et son terminal mobile avec lequel il fait ses déplacements, parler de la mobilité des utilisateurs revient à parler de la mobilité des terminaux mobiles ; ces deux concepts sont interchangeables. Aussi, sachant que les applications et les services utilisés via les WCNs ne peuvent se faire que par le biais des terminaux mobiles, nous pouvons assimiler la mobilité des terminaux mobiles à celle des services et applications en cours sur ces terminaux mobiles. Ainsi, la mobilité des utilisateurs, la mobilité des terminaux mobiles des utilisateurs et la mobilité des services utilisés via les WCNs par le biais des terminaux mobiles peuvent être interchangeables dans les WCNs.

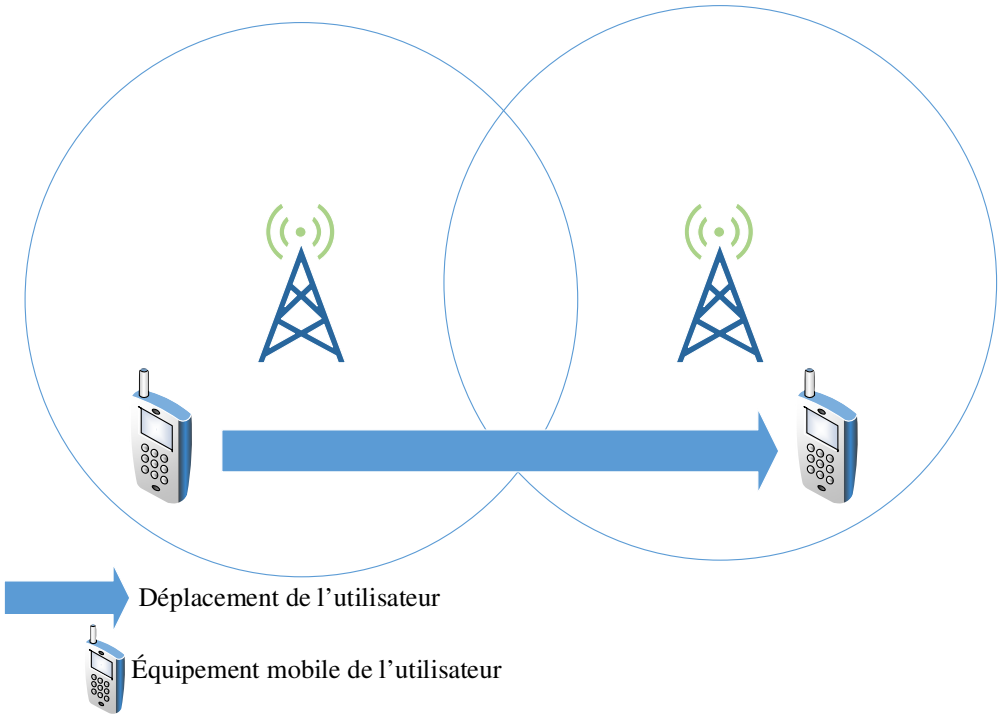


Figure 4 : Illustration d'un transfert intercellulaire

2.2. Les modèles de mobilité et la prédiction de la mobilité des utilisateurs dans les WCNs

Un grand nombre de travaux ont été menés sur les modèles de mobilité et la prédiction de mobilité des humains [16, 20, 22-25, 44-57]. En effet, la prédiction de la mobilité se base sur les modèles de mobilité et est primordiale pour la gestion et la planification dans plusieurs domaines autres que les WCNs tel que le domaine du transport. Vu le caractère mobile des utilisateurs dans les WCNs, la prise en compte de leurs mobilités devient une nécessité pour la gestion efficace des ressources du réseau. Le critère d'évaluation principal des travaux menés sur la prédiction de la mobilité des humains dans les WCNs est la similarité entre les chemins prédits et les chemins utilisés réellement par les humains. Cependant, l'objectif final de la prédiction du chemin devant conduire au support de la QdS, deux autres critères doivent être pris en compte : (1) les chemins doivent être des séquences de routes afin de savoir les points d'entrée et de sortie des cellules ; ces données sont utiles pour l'estimation des temps d'arrivée et de sortie des cellules et (2) la longueur du chemin prédit doit aller au-delà d'une seule

cellule pour permettre un contrôle d'admission sur plusieurs cellules . Dans cette section, nous présenterons brièvement les modèles de mobilité, la gestion de la mobilité dans les WCNs avant d'aborder largement les travaux effectués sur la prédiction de la mobilité.

2.2.1. Les modèles de mobilité

Dans [16, 46, 58, 59], les auteurs se sont intéressés aux modèles de mobilité de façon générale. Dans [59], Bai *et al.* classifient les modèles de mobilité en deux groupes. Premièrement, les modèles basés sur les déplacements aléatoires des utilisateurs ; dans ce groupe de modèles, la destination, la vitesse et la direction des déplacements sont choisies aléatoirement et aussi indépendamment des autres utilisateurs. Par conséquent, ce groupe de modèles ne peut pas être applicable aux utilisateurs mobiles dans les WCNs. Secondairement, les modèles qui prennent en compte l'aspect spatial, temporel et géographique ; l'aspect spatial fait référence aux contraintes liées aux déplacements des autres utilisateurs tandis que le temporel souligne les contraintes dues aux lois de la physique en termes des déplacements. Quant à l'aspect géographique, il renvoie aux contraintes d'ordre physique de la zone où se font les déplacements. Ce second groupe de modèles est plus approprié pour la représentation de la mobilité des utilisateurs de WCNs.

Dans [58], les auteurs présentent trois types de modèles de mobilité selon la zone géographique. Ils identifient les modèles régionaux (très souvent utilisés pour la localisation des utilisateurs), les modèles urbains (très souvent utilisés pour la gestion des ressources dans les réseaux sans fil) et les modèles routiers (très souvent utilisés pour l'étude du comportement des utilisateurs sur les rues). Karamshuk *et al.* [46] présentent l'état de l'art sur la mobilité des humains, à savoir les modèles non aléatoires. Ils soutiennent que les modèles de mobilité doivent prendre en compte la nature des déplacements des humains. Les auteurs classifient les modèles selon trois dimensions : la dimension spatiale, la dimension temporelle et la dimension sociale. La dimension spatiale porte sur le comportement des utilisateurs dans un espace physique ; par exemple, les distances à parcourir pendant leurs déplacements. La dimension temporelle porte sur la variation des caractéristiques des déplacements des utilisateurs avec le temps ; par exemple, la durée du temps passé dans un lieu. La dimension sociale porte sur l'interaction et les affiliations entre les utilisateurs. Ils distinguent également

les modèles de mobilité basés sur l'exploitation de la préférence des lieux [60-62], ceux qui utilisent les relations sociales [63-68] et ceux qui s'appuient sur les activités quotidiennes des utilisateurs [45, 69]. En plus, dépendamment du type de données utilisées pour l'étude du modèle de mobilité, ils distinguent deux classes : une première basée sur l'usage de données historiques des déplacements pour l'extraction des statistiques d'intérêt ; une seconde fondée sur la connaissance des goûts des utilisateurs en termes de préférence de lieux à visiter. Cependant, le type de modèles basé sur l'usage des données historiques (première classe) est fortement lié à la zone où a eu lieu l'étude et ne peut pas être appliqué hors de cette zone.

Dans [16], Vassilya et Isik présentent le modèle markovien, le modèle de groupe et le modèle basé sur les activités. Le modèle de groupe est un modèle généraliste qui ne prend pas en compte l'utilisateur comme une entité individuelle, mais considère l'ensemble des utilisateurs ; ce modèle considère l'ensemble des habitudes des déplacements des utilisateurs pour optimiser l'utilisation du réseau. Cependant, il ne permet pas d'avoir des prévisions sur le mouvement des utilisateurs. Le modèle markovien permet de prédire les futurs déplacements des utilisateurs en se basant sur les déplacements historiques ; ces déplacements sont utilisés pour les calculs de probabilité de transfert entre chaque cellule. Malheureusement, c'est un type de modèle coûteux en calcul. Le modèle basé sur les activités est une extension du modèle markovien ; dans ce type de modèle, les paramètres tels que le temps de la journée, la position courante et la destination prévue sont pris en compte. Il faut noter que ce modèle est plus complexe que le modèle markovien et requiert une plus grande quantité de données.

Dans notre étude, nous fusionnons les modèles urbains et routiers afin de proposer un modèle basé sur les activités. Ce qui permet d'accroitre la précision de notre modèle de prédiction de la mobilité. Nous prenons également en compte les trois dimensions citées par [46]. De même, nous ne nous limitons pas à l'usage de données historiques, mais faisons usage de la connaissance des goûts des utilisateurs en termes de préférence de lieux à visiter. L'objectif des modèles de mobilité dans les WCNs étant de permettre une gestion efficace de la mobilité des utilisateurs afin de supporter la QdS, nous présentons brièvement la gestion de la mobilité dans la section suivante.

2.2.2. Les domaines d'implication de la mobilité dans les WCNs

La gestion de la mobilité des utilisateurs dans les WCNs est subdivisée en deux domaines : la gestion de la localisation et la gestion du transfert intercellulaire. La figure 5 montre les différents domaines de la mobilité dans les WCNs.

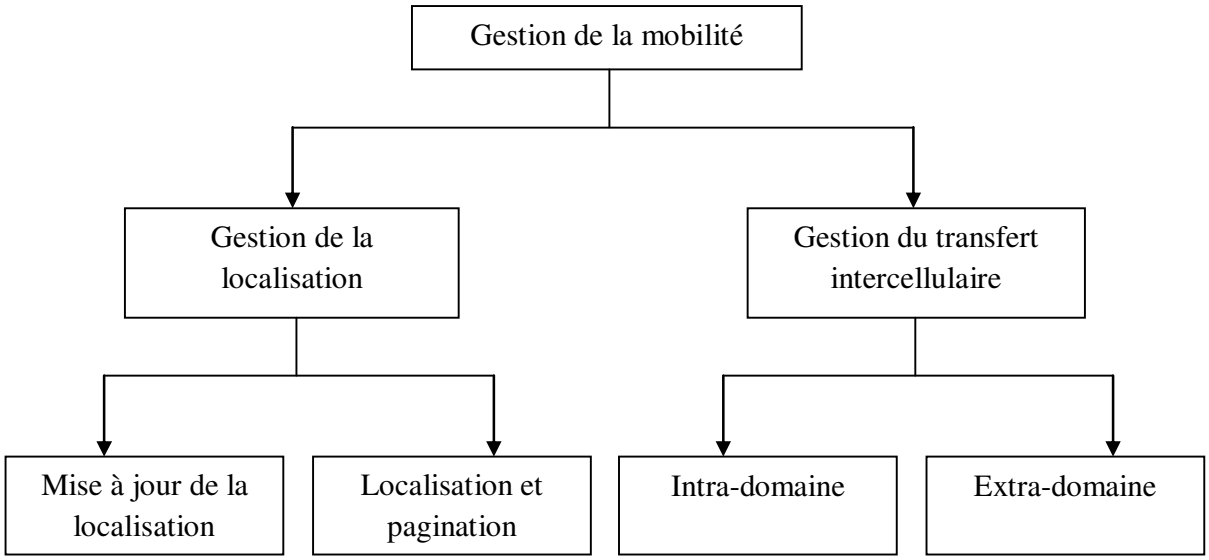


Figure 5 : Différentes composantes de la mobilité dans les WCNs.

La gestion de la localisation [22, 47, 70-72] garde les traces des différentes localisations des utilisateurs mobiles. Elle comprend deux tâches principales : (1) la mise à jour ou l'enregistrement de la localisation. Pour le réaliser, l'utilisateur informe périodiquement l'épine dorsale du réseau de mettre à jour les bases de données de localisation avec ses dernières informations de localisation [43, 73] et (2) la livraison d'appel qui consiste à déterminer la position courante de l'utilisateur par l'épine dorsale du réseau lorsqu'une communication pour cet utilisateur est amorcée. Pour le faire, l'épine dorsale du réseau se base sur l'information disponible dans ses bases de données de localisation. Deux étapes majeures sont appliquées dans la livraison d'appel : (1) la détermination de la base de données servant l'utilisateur appelé et la localisation de la cellule visitée ou le sous-réseau de l'utilisateur appelé ; et (2) la pagination pendant laquelle les messages sont envoyés dans toutes les cellules ou les sous-réseaux au sein de la zone de localisation de l'utilisateur appelé.

Le but principal de la gestion de la localisation est de réduire le coût des mises à jour des localisations des utilisateurs mobiles tandis que la gestion du transfert intercellulaire a

pour but principal de maintenir la connectivité des utilisateurs malgré leurs déplacements. Le déplacement des utilisateurs est la cause des changements fréquents de point d'attachement à l'épine dorsale du réseau (transferts intercellulaires). Les défis majeurs dans la gestion des transferts intercellulaires sont [43] la réduction des coûts de signalisation et d'alimentation et le support de la QdS pendant le processus de transfert intercellulaire. Par support de la QdS, nous entendons : (1) une très faible latence du temps de transfert entre deux cellules du même réseau ou de réseaux différents ; ce temps comprend le délai de traitement des messages de signalisation, le délai de configuration des routes et des ressources, le délai de transformation de format et le délai d'authentification s'il y a lieu ; (2) une réduction des perturbations du trafic de l'utilisateur (en terme de paquets) ; (3) des taux d'échec du transfert et de perte de paquets quasi nuls ; (4) une utilisation efficace des ressources du réseau ; et (5) une évolutivité accrue, une fiabilité et une robustesse des systèmes du réseau. Des études présentant l'état de l'art sur la gestion de la mobilité des utilisateurs dans les WCNs sont accessibles dans [16, 46, 58, 59, 73].

L'un des aspects les plus importants dans la gestion du transfert intercellulaire est la prédiction de la mobilité ; c'est-à-dire savoir à l'avance les positions qui seront visitées par les utilisateurs pendant leurs futurs déplacements. Les modèles de base [16, 20, 22-25, 44-56] sont des modèles utilisant le critère de position, le critère de direction, le critère de segmentation, le critère du temps et le principe de probabilité conditionnelle. Plus spécifiquement, en s'appuyant sur la régularité des déplacements de l'utilisateur, les modèles de base évaluent la distribution de probabilité du prochain déplacement en se référant à la position de départ de l'utilisateur. Ils identifient la position courante de l'utilisateur puis utilisent les données historiques sur la position de départ du déplacement pour prédire le prochain déplacement ; la possibilité de déplacement qui a la plus grande probabilité est choisie comme prochain déplacement ; en d'autres termes, la cellule la plus fréquentée en fonction de la position de départ du déplacement et de la position courant est choisie comme la prochaine à visiter. Le critère sur la direction permet d'étendre le critère de position en incluant la direction de l'utilisateur afin d'accroitre la précision. Le critère de segmentation permet d'étendre le critère de direction. Tous les précédents déplacements sont segmentés et chaque segment est composé d'une cellule de départ et d'une cellule d'arrivée. Certains

modèles de base utilisent donc la correspondance entre les segments enregistrés et la trace du déplacement actuel de l'utilisateur pour effectuer la prédiction de la mobilité des utilisateurs. Le critère du temps permet de prendre en compte le fait que la régularité des déplacements des utilisateurs est liée aux différents moments de la journée.

Ces études sur la prédiction de la mobilité des utilisateurs sont possibles grâce à la disponibilité des données historiques réelles (les traces de déplacements) des utilisateurs. Ces données sont collectées soit par les opérateurs de réseaux [60], soit dans le cadre d'études académiques [74, 75], soit encore par la communauté d'Internet [76]. Il existe deux systèmes de collecte de ces données : le système utilisant la localisation des coordonnées géographiques par satellite (p.ex., le système de localisation globale GPS) et le système utilisant la localisation des points d'accès (p.ex., les stations de base) des WCNs. Cependant, les données les plus précises sont celles qui sont fournies par les systèmes utilisant les données satellites [23, 77-79]. Ce type de données permet de prédire le chemin (séquence de segments de route) qu'un utilisateur pourrait emprunter pour atteindre une destination. Quant aux systèmes utilisant les données sur les points d'accès [22, 60, 80-83], ils se limitent à la prédiction de la séquence de cellules qui seront traversées ; cependant, le fait que les segments de route soient localisés dans les cellules permet de déduire la séquence de cellules à partir de la séquence de segments de route. Il faut noter que, quand bien même les données historiques sur les déplacements des utilisateurs mobiles demeurent le moyen de prédiction le plus utilisée, Tzung-Shi, *et al.* [44] montrent que ce type de données est insensible aux changements de comportement des utilisateurs en termes de mobilité. Effectivement, le changement de comportement d'un utilisateur ne se reflète pas sur la distribution de probabilité de la régularité des déplacements de celui-ci. Lorsque l'utilisateur emménage ou se retrouve dans un nouveau lieu où il n'existe pas de données historiques sur ces déplacements, les modèles de prédiction uniquement basés sur les données historiques deviennent incapables de faire la prédiction. Quant à Song *et al.* [84], ils concluent qu'il ne faut pas se limiter aux données historiques sur les déplacements des utilisateurs, mais qu'il faut prendre en compte l'aspect spatial et temporel.

Dans notre étude, nous proposons un modèle de prédiction des temps de transferts intercellulaires utilisant les données historiques (position, accélération, décélération, vitesse maximale) fournies par des systèmes de localisation par satellite tels que le GPS.

2.2.3. La prédiction de la mobilité des utilisateurs

Dans la littérature, nous rencontrons plusieurs classes de modèles de prédiction de la mobilité selon le nombre de cellules prédites ou la nature des données utilisées. Selon le nombre de cellules prédites, il y a les modèles dont la prédiction se limite à la prochaine cellule (court terme) [24, 26, 47, 49, 85, 86] et ceux qui font la prédiction sur plusieurs cellules (long terme) [20-25]. Toutefois, il est possible d'obtenir une prédiction à long terme en appliquant une répétition de la prédiction à court terme. Selon la nature des données utilisées, il y a des modèles qui utilisent les données historiques sur les déplacements précédents des utilisateurs [20, 22-25, 47, 49, 60, 77-83, 85] et ceux qui n'en font pas usage [29, 67-69, 87]. Cependant, la majorité des modèles proposés utilisent les données historiques, car elles traduisent les habitudes des utilisateurs en termes de déplacements. Malheureusement, ces données ne sont pas toujours disponibles, d'où l'apport de modèles ne s'appuyant pas sur les données historiques.

Afin d'éviter l'utilisation des données historiques sur les déplacements des utilisateurs, certains modèles de prédiction de la mobilité utilisant des approches stochastiques ont été proposés [29, 67-69, 87]. Samaan *et al.* [29] appliquent la théorie de *Dempster-shafer* sur les préférences et les buts des utilisateurs pour prédire leurs mobilités tandis que dans [67, 68], les auteurs utilisent la théorie sociale sur la structure des relations entre les utilisateurs. Ekman *et al.* [69] définissent leur modèle en s'appuyant sur les activités quotidiennes des utilisateurs. Ils assument que la majorité des utilisateurs quittent leurs domiciles pour leurs lieux de travail, leurs lieux de travail pour le restaurant, le restaurant pour leurs lieux de travail, leurs lieux de travail pour un lieu de loisir et enfin de ce lieu de loisir pour leurs domiciles. L'inconvénient principal de ce type de modèle - qui n'utilise pas les données historiques - c'est qu'il se limite à la prédiction de la destination et non du chemin emprunté pour atteindre la destination. La prédiction de la destination sans le chemin vers cette destination ne permet pas d'assurer une gestion efficace des transferts intercellulaires, surtout lorsque la position courante et la

destination sont distantes de plus d'une cellule. De même, l'usage des données historiques pour la prédiction de la mobilité des utilisateurs produit une plus grande précision et ne se limite pas à la prédiction de la destination finale. De ce fait, nous nous focalisons sur les travaux qui font usage des données historiques afin de prédire les futurs déplacements des utilisateurs à court ou long terme. Les travaux faisant usage des données historiques prennent en compte un certain nombre de critères qui sont : l'origine du déplacement, la position précédente, la position courante, la direction du déplacement, la segmentation, la position des voisins, la vitesse courante, la distance du prochain arrêt, la distance de la prochaine cellule, le contexte de l'utilisateur, le comportement collectif des utilisateurs, le jour et le temps. Cependant, la question principale est : quels critères utiliser, comment les utiliser et comment déterminer la meilleure précision avec les critères choisis ?

Wanalertlak *et al.* [49] proposent un modèle de prédiction de la prochaine cellule à visiter (court terme) basé sur plusieurs facettes du comportement des utilisateurs : la localisation (données historiques sur les déplacements), l'identification à un profil (les utilisateurs ayant le même comportement), le moment de la journée (les utilisateurs changent de comportement en fonction du moment de la journée) et la durée qui est catégorisée en « courte », « moyenne » et « longue ». La durée représente le temps que passe un utilisateur dans une cellule ou indirectement, la vitesse d'un utilisateur dans une cellule. En s'appuyant sur les données historiques sur les transferts intercellulaires et la cellule précédente (cellule visitée avant la cellule où l'utilisateur est localisé présentement), les auteurs déduisent une liste des cellules voisines comme étant la prochaine cellule à visiter. Cette liste est ordonnée en tenant compte de la durée de l'utilisateur dans la cellule précédente et son profil. Le temps de la journée est utilisé lorsque le niveau de précision est faible. Malheureusement, leurs modèles ne tiennent pas compte de la structure de l'environnement de déplacement qui est très déterminant pour la prédiction de la mobilité. De même, le niveau de précision peut être élevé malgré une fausse prédiction. Il serait préférable de se focaliser sur une bonne prédiction contrairement à une bonne précision (niveau élevé).

Calabrese *et al* [85] présentent un modèle (court terme) de prédiction de la position d'un utilisateur qui se base sur le comportement individuel et collectif des utilisateurs. Leur modèle considère les déplacements antérieurs (données historiques) des utilisateurs filtrés

selon le jour de la semaine et le temps de la journée. Les auteurs considèrent les caractéristiques géographiques des zones où il y a des mouvements de collectivité en termes d'utilisation des espaces, des points d'intérêt et la distance pour atteindre ces zones. Dans leur proposition, la probabilité qu'une cellule soit la prochaine à être visitée est égale à la fréquence des visites antérieures de cette cellule (en tenant compte de la cellule précédente) sur une période de temps T, à savoir (k - T), (k -2T), etc. Malheureusement, ce modèle ne considère pas la position de départ du déplacement. Faire usage du jour de la semaine comme filtre est une bonne idée, cependant la période considérée (hebdomadaire) demande plusieurs jours pour peu de périodes. De même, le comportement de l'utilisateur peut changer chaque semaine si son emploi du temps est hebdomadaire. Il serait judicieux de considérer le type de jour au lieu du jour de la semaine. Cela permet aussi de réduire le nombre de jours et d'accroitre le nombre de périodes. Calabrese *et al* [85] assument que les lieux à usage collectif les plus proches de la position actuelle de l'utilisateur sont plus probables d'être la destination et effectuent leur prédiction en y tenant compte. Ce qui n'est pas une mauvaise idée, cependant déterminer la destination sur la base du comportement individuel de l'utilisateur donne plus de précision.

Anagnostopoulos *et al.* [47] proposent un modèle de prédiction de la position future des utilisateurs (court terme) basé sur les données historiques des déplacements et sur l'environnement de déplacement de l'utilisateur (par exemple, des cellules voisines). Plus précisément, selon la trajectoire du déplacement actuel (c'est-à-dire l'ensemble ordonné de cellules déjà traversées), les auteurs choisissent les futures trajectoires possibles parmi les trajectoires historiques de l'utilisateur dont le début (les premières positions) est semblable à celui de la trajectoire du déplacement actuel. En d'autres termes, ils font une correspondance entre la trajectoire actuelle et les trajectoires des données historiques de déplacements. Dans cette approche, l'information sur l'environnement de l'utilisateur (les utilisateurs dans les cellules voisines) n'est pas utile et le contexte temporel (à savoir les jours de la semaine et l'heure de la journée) n'est pas exploité.

Byungjin *et al.* [24] proposent un modèle de prédiction à long terme. Faisant usage des données historiques de déplacements des utilisateurs, ils extraient des règles de mobilité. Ces règles sont un ensemble de « tête vers queue » et leurs poids. « Queue » correspond à une

séquence ordonnée de cellules visitées après la visite de « tête » qui est aussi une séquence de cellules ordonnées. Le poids représente le nombre d'utilisations de « tête vers queue ». Lorsqu'un utilisateur approche de la bordure d'une cellule, sa prochaine cellule est prédite en s'appuyant sur les cellules déjà visitées et les règles de mobilité. Le chemin restant vers la destination est la « queue » qui a le plus grand poids. Malheureusement, les données historiques ne sont pas filtrées en fonction des jours de la semaine et des moments de la journée. Aussi, la prise en compte d'une destination estimée aurait réduit les risques d'erreur. Il faut noter que l'extraction des règles de mobilité crée du traitement supplémentaire, car il faut considèrer également les « queues » dont l'utilisateur n'a jamais fait usage.

Mazhelis [23] conçoit un modèle de prédiction de mobilité à long terme basé sur l'apprentissage par l'exemple. Il compare les données historiques des déplacements observés et le trajet actuel de l'utilisateur, puis sélectionne le plus similaire comme futur chemin. Cette approche est semblable à celle de Byungjin *et al.* [24], mais elle nécessite moins de traitements. Les limites du modèle conçu par l'auteur sont le fait qu'il peut fournir plus d'une solution comme futur chemin et qu'il ne filtre pas les données historiques selon les jours de la semaine et les moments de la journée.

Xie *et al.* [25] introduisent un modèle de prédiction de la mobilité à long terme. Leur approche se fonde sur les données historiques des déplacements et les données historiques sur le contexte environnemental des zones où se déplacent les utilisateurs. Le contexte environnemental se compose de l'aménagement urbain, de la météo, du trafic routier, de l'heure, de la date, de la vitesse, de l'air et de la température. En combinant les données historiques et le contexte environnemental, les auteurs extraient des contextes de mobilité. Puis, en effectuant une correspondance entre le contexte de mobilité actuelle et les contextes historiques de mobilité, ils prédisent la mobilité. Un écueil de cette approche est qu'elle nécessite une longue durée de collecte de données (plusieurs saisons ou années). En plus, ce modèle offre plus d'un chemin comme solution.

Jeung *et al.* [20] élaborent un modèle de prédiction de mobilité à long terme basé sur les données historiques de déplacement des utilisateurs. Leur modèle sauvegarde l'historique des virages aux intersections de routes et les vitesses de déplacements sur les segments de routes. À chaque intersection de routes, les auteurs choisissent le segment de route le plus

fréquemment utilisé comme prochain déplacement ; ils prennent en compte le segment de rue précédent. Malheureusement, le modèle proposé ne considère pas l'aspect temporel (jours de la semaine et moments de la journée). En effet, le segment de route le plus utilisé en s'appuyant sur le nombre total de fois que l'utilisateur l'a emprunté peut ne plus l'être si les jours de la semaine et les moments de la journée sont pris en compte. De même, la vitesse ne permet pas de prédire les segments de route qu'un utilisateur pourrait emprunter. Le fait de calculer la probabilité de toutes les possibilités de chemin à chaque intersection de routes augmente la charge de traitement.

En résumé, les modèles de prédiction de la mobilité proposés dans la littérature présentent une ou plusieurs limites.De fait, ces modèles : (1) effectuent la prédiction pour la cellule suivante [26, 27, 47, 49, 70, 75, 85, 86] ; (2) ne considerent pas les types de jours de la semaine et le moment de la journée [20, 24, 47-50] ; (3) ne prennent pas en compte l'origine du déplacement et la destination (estimée ou connue) [20, 22-24, 54, 85] ; (4) proposent plus d'un chemin comme solution [23, 25] ; (5) engendrent un surplus de traitements [20, 24, 47] ; (6) nécessitent plus d'espace de sauvegarde des données [24, 25] ; (7) se basent sur des hypothèses non réalistes (p. ex., les utilisateurs suivent toujours les mêmes chemins [56]) ; ou (8) utilisent uniquement les données historiques [20, 23, 24, 47].

Le Tableau 1 présente un récapitulatif des travaux de recherche sur la prédiction de la mobilité des utilisateurs. Comme souligné, nous nous sommes concentrés sur les travaux utilisant les données historiques. Cependant, certains travaux utilisent d'autres types de données en plus. C'est le cas de Calabrese *et al.* [85] qui utilisent le comportement collectif comme zone d'attraction et Xie *et al.* [25] qui utilisent le contexte environnemental des zones où se déplacent les utilisateurs. Effectivement, l'usage des données de nature variée accroit la précision de la prédiction. Toutefois, ces données sont coûteuses en termes d'espace de stockage et générent des traitements supplémentaires. Nous constatons qu'aucune des approches ne considère la direction vers la destination, qu'elle soit estimée ou supposée, ni la direction courante (vecteur de l'origine/position précédente vers position courante). Ces critères pourraient fortement remédier aux erreurs dues à l'usage des données historiques et autres types de données. Néanmoins, certains travaux utilisent le chemin déjà parcouru entre l'origine et la position courante [23-25, 47]. Une autre solution pour réduire les erreurs des

données historiques est le filtrage selon le type de jour et le moment de la journée. Seuls Calabrese *et al.* [85] utilisent cette idée. Malheureusement, leur méthode demande plusieurs journées de collecte de données. En somme, nous constatons qu'aucune des approches proposées ne fait une combinaison de tous ces critères afin de proposer un modèle de prédiction de la mobilité offrant une précision plus élevée. Un modèle de prédiction de mobilité prenant en compte les limites des travaux existants mentionnées plus haut est présenté dans le chapitre 3.

Tableau 1 : Récapitulatif des travaux sur la prédiction de la mobilité des utilisateurs de réseaux mobiles cellulaires.

Travaux	Données historiques	Autre type de données	Structure spatiale	Destination	Jour de la semaine	Moment de la journée	Direction	Position de départ	Plusieurs solutions	Long terme
Anagnostopoulos *et al.* [47]	√		√					√	√	
Calabrese *et al.* [85]	√	√		√	√	√				
Wanalertlak*, et al.* [49]	√	√				√				
Byungjin *et al.* [24]	√							√		√
Mazhelis [23]	√							√	√	√
Xie *et al.* [25]	√	√	√					√	√	√
Jeung *et al.* [20]	√		√							√
Samaan *et al.* [29]		√				√			√	
Ekman*, et al.* [69]		√				√			√	
Musolesi *et al.* [67]		√							√	
Yuan *et al.* [68]		√							√	

2.3. L'estimation des temps de transferts intercellulaires

Les travaux sur l'estimation des temps de transferts intercellulaires sont utilisés par les systèmes de réservation prédictive de la bande passante. Ainsi, le seul critère d'évaluation de

ces travaux est l'écart entre le temps de transfert intercellulaire estimé et le temps de transfert intercellulaire réel; plus cet écart est petit, plus l'estimation est précise. Cependant, cette estimation des temps de transferts intercellulaires nécessite une forte implication des études dans le domaine du transport. Ce qui rend l'estimation des temps de transferts intercellulaires très complexe; l'objectif final de l'estimation des temps de transferts intercellulaires dans les WCNs étant le contrôle efficace de l'admission des sessions; cette complexité (coût des traitements en temps et en espace mémoire) de l'estimation des temps de transferts intercellulaires est plus élevée car l'estimation du temps de transfert intercellulaire doit se faire sur l'ensemble des cellules qui sont traversées par le chemin de l'utilisateur. Au regard de cette complexité, plusieurs travaux assument plusieurs contraintes (absence de feux tricolores, pas de temps d'arrêt aux intersections de routes ou temps d'arrêt égaux pour tous utilisateurs et vitesse constante sur les routes) afin de contourner la complexité ; les paragraphes suivants présentent certains de ces travaux.

Lu *et al.* [26] proposent un système d'estimation du temps du transfert intercellulaire. Ils se basent sur la distance entre des points de contrôle et la prochaine cellule et la vitesse (constante choisie aléatoirement sur un intervalle prédéfini) de l'utilisateur. Des points de contrôle sont localisés aux intersections entre les routes et les bordures des cellules (limite de la couverture du radio d'une cellule). D'autres points de contrôle sont localisés sur chaque direction possible des intersections entre des routes (à une distance prédéfinie de l'intersection de routes). Les utilisateurs mobiles doivent juste signaler leurs positions à la station de base lorsqu'ils atteignent un point de contrôle. Puis, relativement à la distance entre le dernier point de contrôle traversé par l'utilisateur et la limite de la couverture cellulaire de la prochaine cellule et la vitesse courante de l'utilisateur, les auteurs estiment le temps d'arrivée de l'utilisateur dans cette cellule. L'une des limites principales de cette approche est l'ajout d'équipements supplémentaires qui sont les points de contrôle. De même, les auteurs ne prennent pas en compte les arrêts possibles aux intersections de routes. En plus, leur approche ne permet pas de faire l'estimation sur plusieurs cellules, elle se limite à la prochaine cellule à visiter.

Madhavi *et al.* [27] conçoivent un outil d'estimation du temps du transfert intercellulaire basé sur les informations concernant la localisation de l'utilisateur à deux

époques consécutives. Ils assument que les stations de base possèdent un plan du réseau routier et que les utilisateurs mobiles sont équipés de GPS. Les stations de base sont chargées de l'estimation de la vitesse de l'utilisateur et sa direction de déplacement en se basant sur les localisations GPS de l'utilisateur à deux époques consécutives. Puis en s'appuyant sur la vitesse et le plan du réseau routier, la station de base calcule la probabilité de chacune des cellules voisines d'être visitée et le temps pour l'atteindre. Leur approche utilise une vitesse estimée en se fondant sur le comportement mobile des utilisateurs sur les routes déjà traversées ; rien ne garantit que l'utilisateur aura la même vitesse dans les routes à traverser. Aussi, l'approche proposée ne prend pas en compte les arrêts aux intersections de routes.

Wee-Seng *et al.* [28] introduisent un système d'estimation du temps que pourrait séjourner un utilisateur dans une cellule. Pour cela, les auteurs s'appuient sur les temps mis par les utilisateurs qui ont déjà traversé les différents segments de routes (portion de route entre deux carrefours) localisés dans la cellule que l'utilisateur concerné par l'estimation prévoit traverser. Ils assument que le chemin de l'utilisateur est connu d'avance. Pour chaque segment de route se trouvant dans la cellule concernée, les auteurs calculent la fonction de densité de probabilité du temps que pourrait mettre un utilisateur pour le traverser. Pour le faire, ils considèrent les temps qu'ont mis les utilisateurs qui ont déjà traversé ce segment et calculent la distribution de probabilité. Ensuite, les auteurs utilisent cette distribution de probabilité pour bâtir la fonction de densité de probabilité. Puis, pour avoir la fonction de densité de probabilité du temps mis pour traverser une cellule, les auteurs utilisent la convolution des fonctions de densité de probabilité de chaque segment de route qui est supposé être emprunté par l'utilisateur et se trouvant dans cette cellule. Enfin, en utilisant la fonction inverse de la convolution des fonctions de densité de probabilité et le choix d'un niveau de probabilité souhaité, ils estiment un intervalle de temps ; cet intervalle de temps représente le temps de séjour le plus court et le plus long d'un utilisateur dans une cellule. Le problème de cette approche se trouve au niveau du choix des utilisateurs pour le calcul de la distribution de probabilité. Effectivement, les auteurs ne limitent pas le nombre et ne tiennent pas compte de la zone où se situe l'utilisateur considéré. De même, les temps d'arrêt aux intersections de routes sont aléatoirement choisis sur un intervalle prédéfini. Ce qui ne reflète pas le comportement des utilisateurs et l'état de la circulation.

Meetei *et al.* [18] proposent une estimation de la durée de séjour d'un utilisateur dans une cellule ; ils prennent appui sur les données historiques sur les durées de séjours des utilisateurs dans les cellules qu'ils visitent. Premièrement, les données mobiles des utilisateurs en termes de l'identité de la cellule et la durée du séjour dans cette cellule sont sauvegardées. Puis sur la base des données historiques spécifiques à un utilisateur, les auteurs estiment les cellules sur le chemin de l'utilisateur et le temps qu'il séjournera dans chacune d'elle. L'avantage de cette approche est sa simplicité ; elle n'a pas besoin de prendre en compte l'état du réseau routier ; cependant, elle est moins précise que ceux qui prennent en compte l'état du réseau routier. En plus, l'utilisateur peut bien visiter la cellule, mais ne pas emprunter les mêmes routes. De même, sa vitesse peut varier d'un jour à l'autre de la semaine ou d'un moment à l'autre de la journée ; il serait bien de filtrer ces données historiques selon le type de jours et le moment de la journée. Cela permettra d'accroitre la précision. En outre, malgré le filtrage, il faudrait également prendre en compte l'état actuel de la circulation routière et le comportement des autres utilisateurs sur les routes.

Zhong *et al.* [88] présentent une estimation du temps que prendra un utilisateur pour atteindre sa prochaine cellule. L'approche proposée par les auteurs est semblable à celle de Madhavi *et al.* [27], à savoir utiliser la vitesse et la distance pour calculer le temps pour parcourir cette distance. Cependant, Zhong *et al.* [88] considèrent la vitesse de l'utilisateur lorsqu'il est à une certaine distance de la prochaine cellule au lieu de l'estimer comme le font Madhavi *et al.* [27].

Choi *et al.* [89] proposent une estimation du temps du transfert intercellulaire en utilisant les données historiques sur les durées de séjours des utilisateurs dans les cellules. Leur approche ressemble à une fusion des approches de Wee-Seng, *et al.*[28] et de Meetei *et al.* [18]. Les données sur les durées de séjours des utilisateurs dans les cellules sont sauvegardées, ce que font Meetei *et al*. [18]. Puis, pour une cellule donnée, les auteurs calculent la distribution de probabilité de chaque durée de séjour ; ils s'appuient sur les données historiques filtrées selon le moment de la journée. Ensuite, sur la base de cette distribution de probabilité, les auteurs bâtissent une fonction de densité de probabilité, ce que font Wee-Seng *et al.*[28]. Cependant, Wee-Seng *et al.*[28] n'utilisant pas de données historiques, ils n'ont pas besoin d'opération de filtrage des données. Contrairement à Choi *et*

al. [89], Meetei *et al*. [18] emploient les données historiques du même utilisateur. Ce qui est favorable à la croissance de la précision en associant une opération de filtrage.

Islam *et al.* [90] conçoivent un modèle de prédiction de la mobilité afin d'estimer un intervalle de temps représentant le temps d'arrivée le plus tôt et le plus tard des utilisateurs dans les prochaines cellules. Ils exploitent les paramètres de mobilité tels que la vitesse, la distance et la direction. Les auteurs assument que les utilisateurs sont équipés de GPS et périodiquement transmettent leurs positions à la station de base associée. Pour commencer, les auteurs estiment la distance que pourrait parcourir un utilisateur dans une cellule avant d'atteindre la prochaine cellule. Pour ce faire, ils utilisent la vitesse moyenne de celui-ci dans cette cellule et les données historiques sur son temps d'entrée et de sortie de la même cellule. Puis, en ajustant cette distance estimée, les auteurs obtiennent une distance minimale et une distance maximale. Enfin, sur la base de ces deux distances et de la vitesse moyenne de l'utilisateur, ils obtiennent le temps d'arrivée le plus tôt et le plus tard. Ce modèle ne prend pas en compte le réseau routier de la zone de navigation. Ce qui oblige les auteurs à faire des estimations de distance entre les cellules. En se fondant sur le réseau routier de la zone de navigation, ils obtiendraient les distances entre les cellules grâces aux longueurs des segments de route qui sont statiques ; l'utilisation du réseau routier de la zone de navigation aurait permis d'éviter l'estimation des distances et aurait fourni des données réelles et exactes.

Rashad *et al.* [91] introduisent un système d'estimation du temps de séjour d'un utilisateur dans une cellule en s'appuyant sur les données historiques. Chaque fois qu'un utilisateur arrive dans une nouvelle cellule, son temps d'arrivée et son temps de sortie sont enregistrés. Ce qui permet d'obtenir une base de données historiques contenant les éléments : identificateur (ID) de la cellule, date et temps d'arrivée et date et temps de sortie. Cette base de données est utilisée pour extraire une séquence de cellules qui représente un chemin. Par exemple, considérant un intervalle de temps donné (soit la localisation à chaque 2 minutes) et deux éléments de la base de données historiques (soit [ID cellule 1, date et temps d'arrivée, date et temps de sortie] ; [ID cellule 2, date et temps d'arrivée, date et temps de sortie]), les auteurs procèdent à une opération de transformation qui aboutit au chemin (ID cellule 1, ID cellule 1, ID cellule 1,…, ID cellule 2, ID cellule 2,…, ID cellule 2) dépendamment de la durée du séjour dans chaque cellule (ID cellule 1 et ID cellule 2). Chaque chemin détermine

implicitement la durée du séjour dans les cellules. Enfin, les auteurs calculent la probabilité de chaque chemin et déduisent le temps de séjour dans cette cellule. Cette approche est basée sur les données historiques de séjour des utilisateurs dans les cellules ; ce qui ne permet pas de prendre en compte les évènements actuels qui se produisent sur les routes.

Duan-Shin *et al.* [92] suggèrent une estimation du temps d'arrivée dans la prochaine cellule en se basant sur le plan du réseau routier de la zone de navigation. Tout comme Madhavi *et al.* [27], ils assument que les stations de base possèdent un plan du réseau routier et que les utilisateurs mobiles sont équipés de GPS. Ces auteurs calculent la vitesse de l'utilisateur entre deux positions géographiques mesurées par un GPS en utilisant leur distance de séparation et le temps mis pour la parcourir. Puis, avec cette vitesse et la distance pour atteindre la prochaine cellule, les auteurs estiment le temps d'arrivée à cette cellule.

Le Tableau 2 présente un récapitulatif des travaux de recherche sur l'estimation des temps de transferts intercellulaires des utilisateurs au cours de leurs déplacements. Aucun des travaux ne prend en compte : (1) les feux tricolores ; (2) la densité de la zone de navigation ; (3) le comportement des utilisateurs sur la route (à savoir, la manière de conduire et de s'arrêter au panneau-stop) ; et (4) l'état actuel du trafic routier (congestion, arrêt de la circulation, rue barrée). Pour le temps d'arrêt aux intersections de route, Wee-Seng, *et al.* [28] emploient une valeur aléatoire qui, malheureusement, ne reflète pas le comportement réel des utilisateurs. De même, l'obtention d'une estimation plus précise devrait prendre en compte l'état présent de la circulation routière. Par conséquent, une utilisation des données historiques n'est pas une bonne approche. Notons qu'un modèle est plus utile lorsqu'il permet de prévoir les temps d'arrivée dans une séquence de cellules représentant le chemin d'un utilisateur. Vu sous cet angle (plusieurs cellules sans utilisation de données historiques), seul le travail de Wee-Seng, *et al.*[28] peut être considéré. Cependant, plusieurs éléments doivent être pris en compte afin de le parfaire. Ces éléments sont : la durée d'arrêt aux intersections, les feux tricolores, la densité de la zone de navigation, le comportement des utilisateurs et l'état du trafic routier. Il est également préférable de baser l'estimation des temps de transferts intercellulaires sur la vitesse ou le temps de séjour selon la densité des zones de navigation. Cela permet d'accroitre la précision. En plus, la vitesse doit être considérée comme une fonction du temps au lieu d'une moyenne ou une constante pour respecter les lois de la

physique en termes de mouvement d'un corps. Un modèle d'estimation des temps de transferts intercellulaires prenant en compte toutes ces remarques est présentée dans le chapitre 4.

Tableau 2 : Récapitulatif des travaux de recherche sur l'estimation du temps du transfert intercellulaire des utilisateurs.

Travaux	Plusieurs cellules	Unité zone d'étude	Vitesse	Durée d'arrêt aux intersections	Ajout équipements	Plan du réseau routier	Données historiques	Filtrage des données historiques	Feux tricolores	Densité de la zone de navigation	Comportement des utilisateurs	État actuel du trafic routier
Lu, *et al.* [26]			va		√							
Madhavi *et al.* [27]			dt			√						
Wee-Seng, *et al.*[28]	√	r		√		√						
Meetei *et al.* [18]	√	c					√					
Zhong *et al.* [88]			vc									
Choi *et al.* [89]	√	c					√	√				
Islam et al. [90]	√		vm				√					
Rashad *et al.* [91]	√	c					√	√				
Duan-Shin *et al.* [92]			dt			√						

vc : vitesse courante ; vm : vitesse moyenne ; va : choix aléatoire ; dt : distance sur le temps ; r : route ; c : cellule.

2.4. Le contrôle d'admission et la gestion de la bande passante

La gestion de la bande passante et le contrôle d'admission des sessions (CAC) sont deux concepts étroitement liés ; une bonne gestion de la bande passante est basée sur un contrôle d'admission efficace. Le CAC se réfère à la tâche de décider si oui ou non une demande de connexion (pour une nouvelle session ou une session en cours) peut être acceptée

et supportée par le réseau. L'objectif principal des travaux sur la gestion de la bande passante et du CAC dans les WCNs est de réduire à la fois la congestion et la rupture inopinée et abrupte des sessions en cours entrainant l'arrêt forcé des appels (applications ou services) associés. Afin d'atteindre ses objectifs, le CAC doit tenir en compte de la quantité de bande passante disponible. Toutefois, pour accroitre la satisfaction des utilisateurs, il faut permettre une plus grande acceptation des nouvelles demandes de sessions. Il faut également noter que les utilisateurs ne sont pas les seuls à satisfaire, il y a les fournisseurs de WCNs qu'il faut considérer. Étant des hommes d'affaires, leur satisfaction est l'accroissement, ou au pire, le maintien de leur chiffre d'affaires. Ce qui passe par une augmentation ou un maintien des ventes, donc une augmentation des abonnés (utilisateurs). Alors, la satisfaction des fournisseurs de WCNs passe par l'accroissement du niveau de satisfaction des utilisateurs afin d'attirer la clientèle. En somme, pour satisfaire aussi bien les utilisateurs que les fournisseurs, il faut un système de CAC qui peut réduire fortement le taux d'arrêt forcé des sessions en cours tout en offrant un taux acceptable de refus de nouvelles demandes de sessions et qui maintient un taux d'utilisation maximal de la bande passante; ce qui représente les critères d'évaluation pour des propositions de systèmes de gestion de la bande passante et de CAC. Cependant, avoir un système de CAC tel que décrit précédemment passe par une connaissance de la quantité de bande passante disponible à l'avance avec certitude; vu que cette connaissance n'est pas possible, son estimation s'impose, d'où la prise en compte de la mobilité des utilisateurs.

Dans la littérature, les systèmes de CAC, dans les WCNs, sont classifiés selon plusieurs paramètres (1) le nombre de cellules dans lesquelles le contrôle d'admission est effectué (une seule cellule (*orienté-cellule*) [26, 28, 86, 89, 93-97] ou plusieurs cellules (*distribué*) [27, 98]) ; et (2) le mode de traitement des sessions en cours au moment du transfert intercellulaire. On distingue également deux modes (a) pas de priorité aux sessions en cours ; et (b) priorité aux sessions en cours. Le mode « pas de priorité aux sessions en cours » traite les sessions en cours et les demandes de nouvelles sessions sans distinction. En effet, tant qu'il y a de la bande passante disponible, c'est le système du « premier à demander, premier à être servi » qui est en vigueur. Ainsi, une session en cours ou une demande de nouvelle session est servie à condition qu'il existe de la bande passante disponible dans la cellule. Cependant, le principal inconvénient de ce mode est que la probabilité des arrêts

forcés de sessions en cours est relativement plus élevée que prévu. Ce qui est très problématique pour la QdS du point de vue de l'utilisateur. Les utilisateurs préfèrent se voir refuser de nouvelles demandes de session plutôt que d'être interrompus de façon inopinée pendant une session. Par contre, dans le mode « priorité aux sessions en cours » [26-28, 86, 89, 92, 93, 95, 97-103] les sessions en cours sont prioritaires sur les nouvelles demandes de session. Ce qui conduit à une plus faible probabilité d'arrêt forcé des sessions en cours. Malheureusement, ce mode engendre une probabilité plus accrue de refus de nouvelles demandes de session. Le concept de base de toutes les approches de priorité de sessions en cours est de donner la priorité aux requêtes des sessions en cours sur les demandes des nouvelles sessions ; c'est-à-dire que les besoins en termes de bande passante des sessions en cours sont prioritaires, même si la demande est faite après celles des nouvelles sessions. Ce mode offre une amélioration de la performance en termes de réduction du trafic (en paquets) total admis. Toutefois, l'amélioration de la performance est liée à la façon dont chaque système donne la priorité aux sessions en cours.

Comme nous l'avons dit plus haut, le but du CAC est de permettre une gestion efficace de la bande passante disponible. De même, nous avons vu que le mode de CAC le plus efficace est celui qui donne une priorité aux sessions en cours. De ce fait, nous nous intéressons à ce mode de CAC dans la suite de la revue de littérature sur la gestion de la bande passante. De façon générale, les approches de gestion de la bande passante utilisant le mode de priorité aux sessions en cours effectuent une réservation de la bande passante dans les cellules pour les sessions en cours. Dans la littérature, nous identifions deux formes de gestion de la bande passante qui se basent sur le mode de priorité aux sessions en cours : la forme statique et la forme dynamique. La forme statique réserve une quantité prédéfinie de bande passante dans chaque cellule pour les sessions en cours. Par contre, la forme dynamique réserve une quantité qui varie en fonction de la demande réelle des sessions en cours (ajustée selon les besoins des sessions en cours). La forme dynamique est plus complexe à mettre en place que la forme statique. Cependant, la forme dynamique permet d'accroitre les performances du réseau. Pour l'ajustement dynamique de la quantité de bande passante à réserver aux sessions en cours, la littérature présente deux approches. Ces deux approches se distinguent par le critère qui permet d'effectuer l'ajustement dynamique de la quantité de bande passante à réserver ; il s'agit du comportement des cellules ou de celui des utilisateurs mobiles. Les

approches qui s'appuient sur le comportement des cellules sont dites « orientés-cellule » [14, 27, 94, 97, 98, 102-106]. Quant aux approches qui s'appuient sur le comportement des utilisateurs mobiles, elles sont dites « orientés-mobilité » [27, 28, 89, 93, 95, 96, 100, 101]. Les approches orientés-cellule sont plus simples, mais moins efficaces que les approches orientés-mobilité qui elles sont plus complexes. Effectivement, les approches orientés-mobilité se basent sur la mobilité individuelle de tous les utilisateurs du réseau et effectuent une réservation de bande passante de façon individuelle. Par contre, les approches orientés-cellule effectuent une réservation globale pour l'ensemble des utilisateurs sans faire de distinction de quantité pour chacun d'eux. Les réservations effectuées par les approches orientés-cellule s'appuient sur l'historique de disponibilité de la bande passante dans les cellules. Dans les paragraphes qui suivent, nous présentons en détail certains des travaux les plus récents sur la gestion de la bade passante et de CAC dans les WCNs.

Esmailpour *et al.* [94] élaborent un système *« orienté-cellule »* supportant les trafics (paquets) hétérogènes avec une prise en compte de la QdS dans les réseaux WiMAX. Ils ajustent l'allocation de la bande passante en fonction du comportement du trafic dans le réseau et des états du réseau. Le comportement du trafic est caractérisé par le taux d'arrivée du trafic, tandis que les états du réseau sont associés à deux métriques de performances calculées sur une période de temps. Ces métriques sont : l'équité (le taux de la quantité de bande passante allouée par rapport à la quantité de bande passante requise) et l'utilisation (le taux du débit obtenu par rapport à la bande passante allouée). Selon la valeur estimée des métriques (équité et utilisation) à l'état courant du réseau et la valeur réelle de ces métriques à l'état précédent du réseau, les auteurs calculent le taux d'arrivée du trafic à l'état courant du réseau. Ils effectuent l'éligibilité des nouvelles sessions sur la base de la quantité de bande passante requise, la quantité de bande passante disponible et l'état du réseau. Ils effectuent trois tests : le test de bande passante (si la bande passante demandée est disponible), le test de l'équité et le test de l'utilisation. Si le test de la bande passante est positif, ils effectuent les deux autres tests. Si ceux-ci le sont aussi, la nouvelle demande est acceptée, sinon elle est refusée.

Jun *et al.* [98] proposent un système de CAC basé sur le comportement des cellules. Ils fournissent une cartographie de la bande passante disponible dans les cellules. Plus précisément, les auteurs calculent les quantités moyennes de la bande passante en utilisant les données historiques des observations sur les quantités de la bande passante dans les cellules.

Une nouvelle session sera acceptée si la bande passante disponible estimée est suffisante sur le chemin. Cela veut dire qu'il faut une disponibilité suffisante dans toutes les cellules qui sont supposées être visitées par l'utilisateur demandeur de la nouvelle session. Autrement, la demande de nouvelle session sera refusée. Malheureusement, les observations historiques de la bande passante dans les cellules ne fournissent pas une estimation précise de la bande passante disponible contrairement à la prise en compte de la mobilité individuelle des utilisateurs. De même, aucun mécanisme de secours n'est proposé pour remédier aux erreurs d'estimation de la bande passante disponible.

Jung-Shyr Wu *et al.* [97] soumettent une approche de CAC qui se base sur le système d'inférence floue. Ils ajustent dynamiquement le seuil d'admission en vue d'obtenir une meilleure utilisation de la bande passante. Pour le faire, les auteurs utilisent deux paramètres (1) la charge des cellules en termes de trafic (quantité de bande passante utilisée par les sessions en cours) ; et (2) le taux d'utilisateurs se déplaçant à grande vitesse dans les cellules. Ces deux paramètres représentent les variables d'entrée du système d'inférence floue et la valeur du seuil d'admission représente la variable de sortie. Pour chaque combinaison des valeurs prises par les variables d'entrée, le moteur d'inférence déduit la valeur de sortie la plus appropriée en se basant sur des règles prédéfinies. Puis la station de base utilise le seuil d'admission sélectionné pour effectuer l'admission des nouvelles sessions. Une nouvelle session sera acceptée dans une cellule si la bande passante disponible est suffisante dans cette cellule et que le seuil d'admission sélectionné pour cette cellule est supérieur au taux d'arrivée des nouvelles sessions.

Wee-Seng *et al.* [28] suggèrent une approche de réservation de bande passante pour les sessions en cours qui prend en compte la mobilité des utilisateurs. Chaque cellule réserve une certaine quantité de bande passante, appelée réservation cible. La réservation cible dans une cellule ne peut être utilisée que par les sessions en cours au moment du transfert intercellulaire vers cette cellule. La réservation cible est ajustée régulièrement en fonction des temps estimés des transferts intercellulaires des utilisateurs. Plus précisément, les auteurs estiment le temps d'arrivée des utilisateurs dans les cellules en s'appuyant sur les temps mis par les utilisateurs qui ont déjà parcouru ces cellules. Ensuite, sachant la quantité de la réservation cible, une demande de nouvelle session sera acceptée dans une cellule si la bande passante disponible dans celle cellule à laquelle l'on soustrait la réservation cible pour cette cellule demeure

suffisante. Dans le cas contraire, la demande de nouvelle session sera refusée. Le choix de la population pour le calcul de la distribution de probabilité ne permet pas d'améliorer la précision des temps d'arrivée dans les cellules. De même, le processus d'acceptation des nouvelles demandes de session se limite à la cellule courante alors qu'une nouvelle session acceptée peut être prématurément arrêtée à la prochaine cellule.

Shufeng *et al.* [95] introduisent une approche de gestion de la bande passante offrant la priorité aux sessions en cours basée sur la mobilité des utilisateurs. Pour cela, les auteurs définissent une cellule virtuelle en fonction de l'emplacement de l'utilisateur. Les coordonnées des bordures de cette cellule virtuelle sur l'axe des abscisses sont $[x-D, x+D]$; x est l'abscisse de la position courante de l'utilisateur et D est la distance séparant les centres de deux cellules réelles. Puis, les auteurs comptent le nombre de sessions transférables, c'est-à-dire les sessions des utilisateurs dans la zone d'intersection des deux cellules virtuelles. Une demande de nouvelle session sera acceptée lorsque le nombre de sessions dans la cellule virtuelle moins son nombre de sessions transférables est inférieur à la capacité d'une cellule réelle en termes de nombre de sessions possibles. Dans le cas contraire, la demande de nouvelle session sera refusée. L'approche proposée se limite à la prochaine cellule.

Mokdad *et al.* [101] proposent un CAC dans les WCNs qui donne la priorité aux sessions en cours plutôt qu'aux nouvelles sessions. Lorsqu'une session en cours arrive dans une cellule, elle sera admise s'il y a de la bande passante disponible. Cependant, lorsqu'une nouvelle session est demandée dans une cellule, elle ne sera acceptée que lorsque le nombre de sessions dans cette cellule est inférieur à une valeur prédéfinie N_1. Dans le cas contraire, la demande sera refusée. En d'autres termes, l'approche proposée par les auteurs permet de réserver une quantité statique de bande passante pour les sessions en cours. Ce type d'approche ne permet pas une utilisation efficace de la bande passante disponible. Effectivement, lorsqu'une cellule atteint son niveau d'acceptation de demandes de nouvelles sessions, il est possible qu'une nouvelle session soit acceptée dans cette cellule. Pour cela, il faut que la bande passante réservée pour les sessions en cours à venir (qui doivent arriver plus tard) soit suffisante pour répondre à la demande de la nouvelle session et que le demandeur de cette nouvelle session soit sorti de la cellule avant l'arrivée des sessions en cours à venir.

Huang *et al.* [93] présentent une approche de réservation de la bande passante utilisant la mobilité des utilisateurs. Les auteurs estiment la bande passante à réserver pour un utilisateur dans chacune des cellules voisines en fonction (1) de la distance entre la position courante de l'utilisateur et son point d'entrée dans la nouvelle cellule ; (2) de la quantité de bande passante requise par l'utilisateur ; et (3) la probabilité de l'utilisateur de visiter la nouvelle cellule. Plus spécifiquement, la quantité de bande passante à réserver est égale au produit d'une constante C, de la probabilité de visite de la cellule à la puissance x_1, de la bande passante requise à la puissance x_2 et de l'inverse de la distance après qu'elle soit élevée à la puissance x_3. Les valeurs x_1, x_2 et x_3 sont dépendantes de la probabilité d'arrêt forcé des sessions en cours. Reddy *et al.* [96] proposent aussi une approche semblable à celle de Huang *et al.* [93]. Cependant, Reddy *et al.* [96] calculent les valeurs x_1, x_2 et x_3 en utilisant une technique de régression de vecteur qui permet de minimiser les risques d'erreur de l'estimation. Son problème majeur est que la réservation est effectuée dans toutes les cellules voisines et aussi, la quantité réservée varie avec la localisation de l'utilisateur à chaque mouvement.

Kim [106] conçoit des algorithmes de gestion de la bande passante pour les WCNs basés sur le concept de solution de négociation. Il propose une fonction d'utilité qui se fonde sur la probabilité d'arrêt forcé d'une session en cours et de refus de nouvelle session. L'auteur estime ces probabilités en utilisant les données historiques sur le trafic ; ces données sont la fréquence de demande de nouvelle session, la capacité en bande passante, la taille des canaux, la durée des sessions. Dans cette approche, l'usage des données historiques ne donne pas l'état réel du réseau.

Tsiropoulos *et al.* [102] proposent une approche probabiliste de CAC en fonction de la quantité de bande passante utilisée dans la cellule. L'idée principale de leur contribution est de traiter les demandes de nouvelles sessions avec une faible priorité par rapport aux sessions en cours. Plus spécifiquement, les auteurs ajustent le taux d'acceptation des nouvelles sessions en fonction de la variation du taux de trafic (sessions en cours) dans le réseau. Lorsque le taux de trafic augmente, les sessions à faible priorité sont progressivement arrêtées pour céder la bande passante qu'elles utilisaient aux sessions de forte priorité. En d'autres termes, le nombre d'acceptations de demandes de nouvelles sessions diminue progressivement avec la quantité

de bande passante disponible. Par contre, les sessions hautement prioritaires seront toujours admises tant qu'il y a de la bande passante disponible.

Khanjari *et al.* [103] présentent une approche adaptative pour le CAC selon la priorité des sessions. Ils regroupent les sessions par classe et définissent une valeur maximale et minimale de bande passante requise selon la classe de chaque session. Une nouvelle session de classe prioritaire sera acceptée si la bande passante est disponible pour satisfaire sa valeur maximale de la bande passante requise. Pour les autres classes, une nouvelle session sera acceptée si la bande passante est disponible pour satisfaire sa valeur minimale de la bande passante requise. La même politique est employée pour les sessions en cours lors du transfert intercellulaire. Lorsqu'il arrive une insuffisance de bande passante pour satisfaire la valeur minimale de bande passante requise par des sessions en cours de classe prioritaire, les sessions en cours des classes les moins prioritaires sont interrompues. Cette approche ne donne pas explicitement la priorité aux sessions en cours.

Mallapur *et al.* [14], Son Vo *et al.* [104] et Ravichandran *et al.* [105] introduisent des approches de gestion de la bande passante basées sur le système d'inférence floue, mais avec des paramètres différents du réseau. Ils donnent tous la priorité aux sessions en cours. Les auteurs appliquent la règle « SI-ALORS » de la théorie du flou sous la forme « SI état ALORS conclusion », où état représente la condition dans laquelle se trouve le réseau. État est la valeur d'un paramètre du réseau ou une combinaison logique de plusieurs paramètres du réseau. Les paramètres du réseau considérés sont la charge de trafic (quantité de bande passante utilisée par les sessions en cours), la bande passante disponible, le type de sessions, la priorité affectée aux sessions et le délai de transmission. La valeur numérique d'un paramètre est associée à une valeur linguistique qui peut être faible, moyenne ou élevée. Ainsi, à partir de la combinaison logique des valeurs linguistiques de chaque paramètre du réseau et des règles prédéfinies, les auteurs décident de l'acceptation ou du refus d'une demande de nouvelle session. Les règles étant basées sur des observations, elles ne permettent pas d'offrir une précision sur l'état du réseau. Cette approche, non plus, ne permet pas de prédire la bande passante disponible à un temps précis.

Le Tableau 3 présente un récapitulatif des travaux de recherche sur le CAC et la gestion de la bande passante ; seules les approches proposées par Mokdad *et al.* [101] et Khanjari et al. [103] ne sont pas dynamiques; en plus, l'approche proposée par Khanjari et al.

[103] ne donne pas la priorité aux sessions en cours. Nous remarquons aussi que seules les approches proposées par Jun *et al.* [98] et Kim [106] sont distribuées. En effet, les approches distribuées sont très efficaces pour la réduction des arrêts forcés/inopinés des sessions en cours.

Tableau 3 : Récapitulatif des travaux de recherche sur le contrôle d'admission et la gestion de la bande passante.

Travaux	CAC distribué	Gestion de la bande passante	
		orientée-mobilité	orientée-cellule
Esmailpour *et al.* [94]			√
Jun *et al.* [98]	√		√
Mokdad *et al.* [101]			√
Jung-Shyr *et al.* [97]			√
Wee-*Seng et al.* [28]		√	
Shufeng *et al.* [95]		√	
Huang *et al.* [93]		√	√
Reddy *et al.* [96]		√	√
Tsiropoulos et al. [102]			√
Khanjari et al. [103]			√
Mallapur et al. [14]			√
Son Vo *et al.* [104]			√
Ravichandran *et al.* [105]			√
Kim [106]	√		√

De même, les approches fondées sur la prédiction de la mobilité des utilisateurs (orientées-mobilité) offrent une plus grande précision contrairement aux approches orientée-cellule. Malheureusement, les approches orientées-cellule, très souvent distribuées, utilisent

les données historiques. Ce qui ne représente pas l'état réel du réseau. Selon le Tableau 3, aucune des approches orientées-mobilité ne propose un CAC distribué ; cela est surement dû à la complexité de proposer une approche à la fois distribuée et orientée-mobilité. Aussi, une combinaison de l'approche orientée-mobilité et de l'approche orientée-cellule peut permettre d'accroitre les performances ; la considération du comportement des cellules peut atténuer l'effet néfaste des erreurs de la prédiction de la mobilité. Ainsi, il est intéressant de proposer une approche distribuée à la fois orientée-cellule et orientée-mobilité permettant une réservation dynamique de bande passante pour les sessions en cours. Le système de réservation doit prendre en compte les temps d'arrivée des utilisateurs (qui utilisent les sessions en cours) dans chaque cellule le long du trajet de l'utilisateur ; une telle approche est présentée dans le chapitre 4.

2.5. Les plateformes d'intégration de modèles de prédiction collective pour la gestion de la bande passante.

Plusieurs plateformes d'intégration pour la gestion de la bande passante dans les WCNs ont été proposées dans la littérature. Cependant, la majeure partie des travaux sur la réservation de bande passante se basent sur le comportement des cellules [17, 98]. Les éléments caractéristiques du comportement des cellules sont (1) les variations de charge (quantité de bande passante utilisée par les sessions en cours) dans les cellules ; (2) les échanges de charge entre cellules voisines ; et (3) l'historique des disponibilités de la bande passante. Malheureusement, les modèles basés sur le comportement des cellules ne donnent pas une meilleure estimation de la bande passante disponible en fonction de la mobilité des utilisateurs. L'exploitation de la prédiction de la mobilité des utilisateurs (comportement mobile des utilisateurs) offre une meilleure estimation de la bande passante disponible dans les cellules. Toutefois, les approches exploitant le comportement mobile des utilisateurs demeurent plus complexes. De même, la plupart des modèles exploitant le comportement mobile des utilisateurs pour l'estimation de la bande passante disponible [27, 86, 101, 107-113] effectuent des prédictions individuelles de mobilité. Ce qui représente une limite pour l'évolutivité de ces modèles. Également, plusieurs des modèles exploitant le comportement mobile des utilisateurs nécessitent des équipements additionnels [26, 27] et ne prennent pas en

compte l'ensemble du chemin (cellules à traversées) de la position courante à la destination [26-28, 86, 91, 97, 114]. Dans les paragraphes ci-dessous, nous présentons brièvement les travaux représentatifs d'intégration de modèles de la prédiction globale/collective de la mobilité des utilisateurs avec des modèles de prédiction de la bande passante disponible dans l'optique d'une gestion efficace de la bande passante [86, 91, 114]. Le critère d'évaluation de ces travaux demeure le taux de rupture de sessions en cours ; toutefois, vu que les modèles de prédiction globale/collective de la mobilité des utilisateurs offre moins de précision que les modèles de prédiction individuelle de la mobilité des utilisateurs, un autre critère à prendre en compte est le coût (délai et énergie) du processus de prédiction de la mobilité. Ainsi, l'objectif de ces travaux est de produire un taux relativement faible de ruptures des sessions en utilisant une prédiction de la mobilité moins coûteuse en traitements. Dans ces travaux, une session en cours sera admise dans la prochaine cellule s'il y a suffisamment de bande passante pour répondre à son besoin en bande passante ; sinon, elle est arrêtée.

K. Dias*, et al.* [86] présentent une plateforme pour le CAC et la réservation collective de bande passante pour les sessions en cours. En s'appuyant sur la localisation GPS d'un utilisateur, ils déterminent la prochaine cellule de celui-ci. Puis en utilisant l'estimation des échanges périodiques de charge entre les cellules voisines, les auteurs prédisent la quantité de bande passante dans les cellules. Ainsi, ils effectuent une réservation virtuelle de bande passante pour les sessions attendues. Une demande de nouvelle session sera acceptée s'il y a suffisamment de bande passante après les réservations virtuelles. Cependant, ce modèle présente quatre limites clés : (1) le processus d'estimation collective se limite à la bande passante disponible et est basé sur le comportement des cellules et non sur la mobilité des utilisateurs ; (2) le processus de prédiction de la mobilité est exécuté pour chaque utilisateur ; (3) l'historique des échanges de charge ne permet pas d'obtenir une meilleure estimation de la bande passante disponible comparée à l'utilisation du comportement mobile des utilisateurs ; et (4) le CAC s'effectue pour une seule cellule ; ce qui n'empêche pas l'arrêt inopiné des sessions en cours après cette cellule.

S. Rashad, *et al.* [91] s'intéressent au problème de la réservation de bande passante et du CAC. Avec la prise en compte de l'analyse des déplacements précédents des utilisateurs, ils génèrent des profils d'utilisateurs. Ces profils d'utilisateurs sont utilisés pour la prédiction des

futurs chemins et des temps de séjour dans les cellules. Le profil est spécifique à une cellule et se base sur les temps de séjour dans les cellules voisines. En effet, selon les temps de séjour dans les cellules, les utilisateurs sont regroupés dans un même profil. Les profils sont associés à une cellule précédente A (avant la cellule considérée B) et une prochaine cellule C (après la cellule considérée B) pour les utilisateurs qui seraient passés dans la cellule A et se trouveraient dans la cellule B ; ce qui forme des chemins-de-profil. Alors, en se fondant sur les chemins-de-profil et la précédente cellule d'un utilisateur ainsi que son profil, les auteurs déterminent, pour l'utilisateur concerné, la prochaine cellule et le temps d'arrivée dans celle-ci. En regroupant les temps d'arrivée dans une cellule des sessions en cours, ils effectuent la réservation de bande passante. Une demande de nouvelle session sera acceptée s'il y a suffisamment de bande passante après les réservations pour satisfaire les besoins de celle-ci. Contrairement à [86], l'estimation de la bande passante disponible s'appuie sur les profils des utilisateurs (comportement mobile des utilisateurs). Cependant, tout comme [86], le CAC se limite aux cellules courantes (pour l'acceptation des demandes de nouvelles sessions) puis aux prochaines cellules (pour la réservation de bande passante pour les sessions en cours).

C.-F. Wu, *et al.* [114] introduisent un modèle de prédiction de l'utilisation de la bande passante et de la probabilité d'arrêt des sessions en cours. Plus spécifiquement, en s'appuyant sur les traces des déplacements antérieurs, les auteurs détectent la périodicité des chemins utilisés et les temps de transferts intercellulaires. Alors, les utilisateurs qui ont les mêmes périodicités sont regroupés dans le même profil de mobilité qu'ils utilisent pour prédire le chemin et les temps de transferts intercellulaires. En effet, si le début d'une occurrence des chemins (séquence de cellules) enregistrés est semblable au chemin courant, les auteurs supposent que le déplacement considéré pourrait être une répétition. Ils choisissent alors la prochaine cellule en se référant à cette occurrence. Cependant, dans le cas de plusieurs occurrences dont les débuts sont semblables au chemin courant, ils sélectionnent celle qui a la plus grande probabilité ; c'est-à-dire le plus grand nombre d'occurrences. Puis, en s'appuyant sur la quantité de bande passante utilisée dans les cellules voisines de la prochaine cellule, les auteurs déduisent la quantité de bande passante dont elle dispose et décide de l'acceptation ou non d'une demande de nouvelle session ; une alerte indique si la nouvelle session pourrait être arrêtée dans la prochaine cellule en cas de transfert intercellulaire. Comme dans [91],

l'estimation de la bande passante disponible se base sur le comportement mobile des utilisateurs. Toutefois, la principale limite de cette proposition est le fait que l'acceptation des demandes de nouvelles sessions se limite à la cellule courante et à la suivante ; contrairement à [86, 91], les auteurs considèrent la prochaine cellule pour l'acceptation des demandes de nouvelles sessions.

Le Tableau 4 présente un récapitulatif des travaux de recherche sur l'intégration des modèles de prédiction collective de la mobilité avec les modèles d'estimation de la bande passante disponible ; les trois travaux proposent une estimation dynamique de la bande passante. Nous remarquons que les modèles de prédiction de mobilité proposés par S. Rashad, *et al.* [91] et C.-F. Wu, *et al.* [114] sont collectifs. Aussi, seuls C.-F. Wu, *et al.* [114] proposent un CAC distribué (sur plusieurs cellules) ; malheureusement, ils utilisent une approche d'estimation de la bande passante disponible orienté-cellule. Seul S. Rashad, *et al.* [91] proposent une approche d'estimation de la bande passante orientée-mobilité. Dans l'ensemble, aucun des travaux présentés est à la fois distribué, orienté cellule et mobilité avec un modèle de prédiction collective de la mobilité ; une telle approche sera présentée dans le chapitre 5.

Tableau 4 : Récapitulatif de quelques travaux de recherche sur l'intégration des modèles de prédiction collective de la mobilité avec les modèles d'estimation de la bande passante disponible.

Travaux	Estimation de la bande passante		Prédiction collective de la mobilité	Distribué
	Orienté-mobilité	Orienté-cellule		
K. Dias*, et al.* [86]		√		
S. Rashad, *et al.* [91]	√		√	
C.-F. Wu, *et al.* [114]		√	√	√

Chapitre 3 :

A Destination & Mobility Path Prediction Scheme for Mobile Networks

Apollinaire Nadembéga, Abdelhakim Hafid, Tarik Taleb

Abstract

Mobile multimedia services are gaining great momentum among subscribers of mobile networks. An understanding of the network traffic behavior is essential in the evolution of today's mobile networks (MNs) and thus leads to a more efficient planning and management of the network's scarce bandwidth resources. The communication efficiency can be largely improved (i.e., optimizing the allocation of the network's limited resources and sustaining a desirable quality-of-service (QoS), if the network anticipates the needs of its users on the move and thus performs reservation of radio resources at cells along the path to destination. In this vein, we propose a mobility prediction scheme for MNs; more specifically, we first apply probability and Dempster-shafer processes for predicting the likelihoods of the next destination, for an arbitrary user in a mobile network, based on user's habits (e.g., frequently visited locations). Then, at each road junction, we apply second-order Markov Chain process for predicting the likelihoods of the next road segment transition, given the path from the trip origin to that specific road junction and the direction to the destination. We evaluate our proposed scheme using real-life mobility traces; the simulation results show that the proposed scheme outperforms traditional schemes.

Index Terms — QoS, mobility prediction, path prediction, destination prediction, mobility model, mobility pattern, and cellular network.

Status: This article is submitted to IEEE Transactions on Vehicular Technology 2013; it is based on the following published papers:

1. A. Nadembéga, A. Hafid and T. Taleb. *A Destination Prediction Model based on Historical Data, Contextual Knowledge and Spatial Conceptual Maps,* IEEE ICC, Ottawa, Ontario, Canada, Jun. 2012.
2. A. Nadembéga, A. Hafid and T. Taleb. *A Path Prediction Model to Support Mobile Multimedia Streaming*, IEEE ICC, Ottawa, Ontario, Canada, Jun. 2012.

3.1. Introduction

Cellular networks have become pervasive in our society and offer high data transfer rate. These characteristics allow mobile users, with portable devices to use various services/applications, such as multimedia streaming, with (Quality of Service) QoS requirements. QoS support in mobile environments is highly challenging because of mobility and resources scarcity [44]. Ensuring QoS, anywhere and anytime, can only be achieved if we are able to predict where a user would likely demand network usage.

Mobility is an inherent characteristic of users in Mobile Networks (MNs); it introduces considerable overhead in mobility management and forwarding services to ensure communication reliability [22]. Indeed, mobile users frequently change their points of attachment to the network; this action is called handoff and is a key element in MNs to provide QoS to the users and support users' mobility [17-19]. During handoff, a user may experience different data streaming rates due to disparity in the bandwidth availability at the different visited cells along the movement path of the user. Frequent changes in streaming rates, mainly those with high magnitude, may severely impact the perceived QoS [21, 30, 107]. Thus, the key current challenge in MNs is to provide a minimum acceptable QoS in each cell visited by the user; this requires prior knowledge of the user's long term movement within a time period d_t (e.g., 30 minutes in advance). Indeed, if the network can predict the user's path (i.e., subsequent transitions of road segments/portions toward destination), it can then provide the user with the required QoS during the whole session. More specifically, the

network will accept the user into the network (e.g., for a multimedia streaming session) only if there are sufficient resources (e.g., bandwidth) in each cell (during the user's residence in the cell) along the predicted path; otherwise, the user will not be allowed to access the network. Thus, a mobility model (more specifically path prediction [16, 20, 22, 24, 46, 47, 50, 60, 68, 70, 71, 75, 115-123]) with reasonable accuracy is of vital importance to provide QoS for mobile users. In this paper, we propose a relatively accurate mobility prediction scheme, called DAMP (Destination And Mobility path Prediction), that allows predicting final or intermediate destinations (e.g., within a time period) and subsequently mobility paths of mobile users (e.g., vehicles, cyclists and pedestrians) based on (1) the trip origin and current location; (2) current and future directions of the mobile users; (3) the current and history of the trajectories followed by the users; and (4) the information on the users' contextual knowledge. DAMP consists of two models, namely Destination Prediction Model (DPM) and Path Prediction Model (PPM).

The objective of DPM is to estimate the user's destination within a time period; it takes into account (a) user's habits in terms of frequently visited locations; (b) direction from movement origin to current location; and (c) user's contextual knowledge. Indeed, making use of the direction from movement origin to current location, DPM determines potential future destinations; accuracy is improved using historical and contextual knowledge. It is possible that a group of potential future destinations may be reached after using the same road segments within a travel time period; thus, DPM performs clustering of possible destinations that aims to reduce the number of potential future destinations; for example, ten potential future destinations can be regrouped into three destination clusters. To form a destination cluster, DPM combines two types of methods: (a) DPM computes, based on second-order Markov Chain, the probability that each possible destination cluster is the next destination cluster making use of filtered historical movement pattern; the filtering process is based on the day and the time of the day to increase accuracy; and (b) DPM builds on the work proposed in [29], wherein mobility prediction is based on evidential reasoning of Dempster-Shafer's theory making use of the user's contextual knowledge. DPM gives different weights to each method according to the number of days in the considered historical movement pattern.

Based on the computed destination cluster, we propose PPM that aims to estimate the path (i.e., subsequent transitions of road segments towards destination) a user would take during his movement from current location towards destination within the time period d_t. PPM takes into account (a) user's habits in terms of the frequency of using road segments to reach a specific destination (e.g., estimated destination cluster); (b) direction from current location to that specific destination; (c) the current trajectory/path (i.e., sub-sequence transitions of road segments from movement origin to the current location); and (d) spatial conceptual maps. More specifically, at each road junction/intersection, PPM determines the next road segment a user will likely use during his movement towards destination; indeed, PPM selects the potential next road segments among the adjacent road segments of the considered/current road junction according to the current direction (i.e., direction from the last crossed road junction to the current road junction) deviation compared with the estimated destination cluster. Then, making use of filtered historical movement trace, PPM computes, based on an extended second-order Markov Chain, the probabilities of all selected potential next road segments given current trajectory/path and destination; the road segment among the potential next road segments with the highest probability is then selected as the next road segment. Here, the historical movement trace filtering process is based on the day of the week (e.g., weekend, holiday and Labor Day) and the time of the day (e.g., morning, noon, afternoon and night). PPM repeats the same process until the selection of last road segment to the destination cluster. The predicted user's path consists of the list of the selected road segments. To the best of our knowledge, this is the first work which takes into account both user's habits and user's contextual knowledge to estimate user's path to destination. In addition, this work is the first to consider destinations clustering to reduce errors using historical and contextual knowledge. More importantly, this work presents one of the few schemes that allow for predicting, without restrictive assumption (e.g., known specific user pattern), the whole path from origin to destination.

The remainder of this paper is organized as follows. Section 3.2 presents some related work. Section 3.3 describes data collection algorithm and database structure, and presents our proposed mobility prediction scheme using second-order Markov Chain. Section 3.4 evaluates the proposed mobility prediction scheme via simulations. Section 3.5 concludes the paper.

3.2. Related Work

Mobility modeling has been extensively studied in many types of wireless networks during the past ten years [16, 20, 22, 24, 46, 47, 50, 60, 68, 70, 71, 75, 115-123]. Mobility model analysis can be used to create models for predicting user mobility. User mobility prediction allows estimating/predicting the location and trajectory of the user in the future. The commonly used mobility models are random walk, random waypoint, fluid flow, Markovian and activity-based mobility models. The simplest of these models are the random walk and random waypoint models; they were originally proposed to emulate the unpredictable mobility of particles in physics. The other models are used for prediction, such as path prediction. It has been shown that users follow daily routines and that mobility models have cyclic properties [16, 20, 22-25, 44-50]. Many researchers rely on such principles to define user mobility prediction models that benefit from the periodic nature of mobility. One of the important fields of users' mobility prediction models is the individual mobility prediction models; basic models [16, 20, 22, 24, 44, 46, 47, 50] are models that employ location, direction, time and conditional probability. Indeed, based on the regularity of user mobility, a conditional probability distribution of next moves is defined considering movement direction and time; the move with the highest value is predicted as the next move. In other words, the cell that was most frequently visited according to the current location, current movement direction and the time of the day is predicted as the next cell.

Recent years have seen a considerable amount of work done on developing users' mobility prediction models. Many of these models [20, 22, 23, 47, 50, 60, 74, 75] heavily rely on the availability of prior information on the users' mobility history. Whereas the continuous tracking of mobile users may lead to better predictions in terms of movement, such models suffer from the large overhead accrued due to constant monitoring; obviously, this requires a more detailed analysis of the users' mobility history, and the application of advanced data mining and knowledge discovery techniques. The models presented in [29, 68, 69, 87] are examples in which the prediction requires no knowledge of the users' mobility history; unfortunately, these models are limited to predicting only where a user is likely to move (i.e., user's final destination) instead of the path to reach this final destination. For example, a scheme that incorporates geographic maps with identifiable landmark objects (e.g., schools,

malls, gyms, libraries) into the users' mobility prediction models has been proposed in [29]; more specifically, the mobility prediction architecture (MPA) [29] gathers the necessary information for the prediction process and analyzes this information using Dempster-Shafer's theory in order to predict future locations of the mobile user. The process of the location prediction is carried out in three main phases: information gathering, evidence extraction and decision making. The information gathering is concerned with capturing the necessary contextual information that includes environment context (places of interest for users and road segments) and user context (user's interests, user's tasks and goals and user's schedule). Making use of this contextual information, they generate bodies of evidence applying the concepts of Dempster-Shafer's theory; for example, the evidence suggesting that a student with a high interest in exercising would be going to the gym is eliminated if his schedule does not allow enough spare time. Then, they compute the belief mass of each body of evidence (i.e., the probability that the body of evidence occurs). Finally, to determine the user's future predicted location, they combine each pair of hypothesis-belief mass using the Dempster rule of combination; a hypothesis represents a location or a sequence of locations. Indeed, they compute the belief function Bel(Hi) of hypothesis Hi. Bel(Hi) describes quantitatively all the reasons to believe in hypothesis Hi. The location with the highest belief value is the predicted future location of the user. However, this technique (i.e., destination prediction model) is used in certain path prediction models [21, 30] to improve prediction accuracy by eliminating or affirming certain paths according to the predicted destination. However, such models require a vast amount of information (e.g., user's preferences, user's goals and user's schedules) to be collected and processed and may not perform very well with temporary changes of the surrounding infrastructure.

A few models consider using both mobility historical data and current conditions in the network. One example is the model proposed in [50], which considers both current trajectory of mobile users (i.e., ordered set of cells already transited) and time-of-day, as well as historical data, to predict the likelihoods of single-cell transition and N-cells transition for an arbitrary user in wireless networks; the prediction model performed quite well for lower values of N (e.g., 2 cells). In [47] a short-term prediction model that employs mobility history for predicting future location of a mobile user while considering the mobile user's current

trajectory within the predefined navigation zone was proposed; this model is limited to only next cell prediction. In [20], a long-term mobility prediction model which considers both current trajectory and movement direction, as well as historical data, was proposed; however, the model requires an immense amount of mobility history and a massive processing load.

In the following, we briefly overview our previous contributions [21, 30] that are most related to the proposed model. In [30], we proposed a method to estimate a user's future destination based on the use of filtered user's mobility history and contextual knowledge; the filter is based on the type-of-day (e.g., working, holiday and weekend) and the time-of-day (morning, noon, afternoon, evening and busy hours); the proposed model also takes into account the movement direction. The drawbacks of this model are (1) the automatic identification of Frequently Visited Locations (FVLs) has not been taken into account; (2) the databases are not updated according to user's predictability level; predictability level is the degree to which a correct prediction of a user's mobility can be made; this degree is related to the frequency of visited places and transited roads according to the day of week and the time of the day; (3) the frequency function of FVL is not explicitly defined, and (4) the weighted sum of the belief and probability functions is not related to user's predictability level. In [21], we proposed an approach which predicts the path the user will use within a time period during his movement from trip origin to destination; the approach makes use of filtered users' mobility history, current movement data (e.g., trip origin and current location) and spatial conceptual maps while assuming a priori knowledge of the destination [30]. More specifically, at each road junction (starting from the location where user first accesses the network), the next road junction the user will likely use during his movement towards the trip destination, is determined. The drawbacks of this approach [21] are (1) the deviation function takes into account the trip origin instead of the previous road junction; and (2) common conditional probability is used instead of second-order Markov Chain that is more appropriate in this type of situations.

To conclude, we summarize the limitations of existing mobility prediction models as follows: (1) they are limited to short term (e.g., next cell) mobility prediction [49, 75]; (2) they do not consider the temporal context (e.g., day-of-week and time-of-day) [20, 24, 47, 50] and/or the whole path from trip origin to current location and direction to destination [20, 22,

24, 25, 47, 50]; these parameters play a key role in improving prediction accuracy; (3) they compute more than on predicted path [25]; (4) they incur high processing overhead [20, 24, 47]; (5) they require massive data storage space [24, 25]; (6) they make restrictive assumptions (e.g., user's movements follow a specific pattern [56]); and (7) they solely rely on the history of individual users' movement [60, 75].

In this paper, we propose a model that proposes solutions to overcome these limitations. The proposed DAMP scheme considers several criteria, namely trip origin to current location, current and future directions, user contextual knowledge, day-of-week and time-of-day in predicting paths; it is a long-term users' mobility prediction model. To limit the impact of using user mobility history (that may change), DAMP considers user knowledge, and regular spatial and temporal patterns for predicting the mobility of users.

3.3. DAMP: DESTINATION AND MOBILITY PATH PREDICTION MODEL

In this section, we present the details of the proposed scheme, called DAMP. More specifically, we present (1) the process used by DAMP to collect data of interest for the prediction procedure, and the structure of the database used to store these data; (2) the semi-Markov process used by DAMP to derive DPM and PPM; (3) the destination prediction model (DPM) that estimates the mobile user destination; and (4) the path prediction model (PPM) that predicts, given the destination computed by DPM, the path the user would take during his movement from current location to destination.

3.3.1. User mobility patterns

3.3.1.1. Assumptions

In this work, we assume that the road topology consists of several roads and junctions while the entire network space is assumed to be divided into cells. We refer to the location frequently visited by a user (e.g., home, school, shop, mall and office) as a frequently visited location (FVL). We assume that a road junction or a FVL 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); we refer to data about visited nodes (e.g., time, date and node ID) as mobility data. We refer to the road between two nodes a and b as a road segment, and identify it using a road segment ID that is represented by the node pair (a, b) where $a \rightarrow b \neq b \rightarrow a$. A user's location is identified by his geographic coordinates. The movement of a mobile user through the network can be described by a list that represents the sequence of road segments that was visited by the user throughout the trip. A user's mobility pattern from the network's perspective is determined by the user's terminal (e.g., mobile phone) mobility pattern. The users' mobility history patterns can be periodically recorded using node ID (road junction and FVL).

The mobility history can either be recorded for each user or collectively for all users into a single history profile per location. The latter method is more suitable for situations where all users generally exhibit similar behavior at a given navigation zone and are also not significantly impacted by erratic behaviors from one or more users. Even though different groups of users have different mobility patterns, it can be difficult to address every type of group behavior in a single mobility model. To derive DPM we need contextual knowledge about users; we assume that User Contextual (UC) information is organized into six categories as shown in Table 5. The UC database can be built (1) by having users fill in a questionnaire and explicitly express their interests with regard to different places within their living areas; or (2) by having users "continuously" registering both their tasks and scheduled appointments. To implement mobility data collection, we assume that (1) user equipment (UE) maintains a database which records data about the user movements and his living area; (2) static data about geographic maps (topology/map of roads), called Navigation Map (NM), is readily available; and (3) UE embeds technology, such as tachometer and GPS, that samples user velocity and coordinates of places visited by the user, along with the day and the time of the visits. It is also assumed that NM database contains geographic coordinates of nodes (e.g., road junctions and FVLs). A FVL is extracted from UC database shown in Table 5 or inserted automatically. Indeed, when a user's velocity is 0 and the current location is not a road segment or road junction, we assume that the current location may be a new visited place and insert it in FVL database. User Movement Trace (UMT) database contains user ID, date *d*, time *t* and node ID (a FVL or a road junction) that represents user location at date *d* and time *t*.

User Frequently Visited Location Trace (UFVLT) database contains user ID, date d, arrival time t_a, departure time t_d and node ID (a FVL) that represents user location at date d from arrival time t_a to departure time t_d; in other words, arrival time t_a denotes the time when the user reaches the location while departure time denotes the time when she leaves it.

In order to limit the size of UMT and UFVLT databases, each entry/record in UMT and UFVLT databases is deleted after a certain number of days, called record lifetime (RL), that is closely related to user's predictability level. Effectively, each user has a predictability level stored in UC; RL decreases when the predictability level increases; a "high predictability" level means the user is more predictable (i.e., it is easier to predict user's movements). For example, for "high (resp. intermediate/low) predictability" level, the user's RL can be set to 2 (resp. 3/4) weeks. Static databases Node, Road junction, FVL, User, UC and NM are updated every six months and also whenever an update becomes required.

Table 5: User Contextual (UC) information structure.

Personal Context (PC)	**Frequently Visited Location Context**	**Task Context (TC)**	**Interest Context (IC)**	**Calendar Context (CC)**	**Day Type (DT)**
•User ID •Name •Age •Predictability	•Location (NM_node) •Location name •Preferableday •Earliestpreferable time •Latestpreferable time •Duration •Characteristics •Importance •Frequency	•Taskname •Earliest time •Deadline •Duration •Characteristics •Importance •Frequency	•Interestname •Preferableday •Earliestpreferable time •Latestpreferable time •Duration •Characteristics •Importance •Frequency	•Location (NM_node) •Date •Time •Characteristics	•From (date) •To (date) •Workday •No workday

The following subsection presents in details the data collection process.

3.3.1.2. Data collection process

Algorithm 1 presents the pseudo-code, executed by UE, for recording data. At every time unit *t* (say 1 s), current velocity of the user and geographic coordinates (*latitude and longitude*) of his current location are measured. When the user's current location is a location (a road segment or a road junction) which is stored in NM database (Line 3), it is inserted together with the current timestamp (date and time) into UMT database (Line 4). When the user's velocity decreases and falls to 0, the user is deemed not moving (Line 5). Thus, his current location is inserted together with the current timestamp (current date and current time) into UFVLT database (Lines 6) when the current location is not a road junction or a road segment (Line 5); it is worth noting that a road junction or a road segment visited by a user cannot be designated as a FVL. The attribute "current time" of current timestamp is assigned to the field "arrival time" in UFVLT database (Line 6). Indeed, arrival time corresponds to the time when velocity falls to 0 while departure time corresponds to the time when velocity starts increasing; in this case, current velocity is different from 0 (Line 8).

Algorithm **1**: **Pseudo-code for movement data gathering.**

Input : User_ID, NM, FVL

Variable: X (Boolean with initial value=true)

Output : UMT, UFVLT and FVL

```
1.  each t sec {
2.      Measure current_velocity and current_location
3.      If (current_location is road junction or road segment)
4.          Put in UMT the 4-tuple (user_ID, current_date, current_time,
    current_location_ID)
5.      If [(current_velocity= 0) and (current_location is not road junction or road
    segment) and ( X=true)] {
6.          Put in UFVLT the 5-tuple (user_ID, current_date, current_time, 0,
    current_location_ID)
7.          X=false
8.      }elseif [(current_velocity≠0) and (current_location is not road junction or
    road segment) and ( X=false)] {
9.          Update the last record of UFVLT set departure_time=current_time
10.         X=true
11.         If (current_location does not exist in FVL)
12.             Put in FVL the 1-tuple (current_location_ID)
13.     }
14.  }
```

Therefore, when the current velocity is different from 0 and the current location is not a road junction or a road segment (Line 8), departure time of the last inserted location in UFVLT database is updated making use of the current time (Line 9). When the last inserted location in UFVLT database (i.e., current location) does not exist in FVL database (Line 11), it is inserted into FVL database (Line 12). The time complexity of Algorithm 1 is $O(n)$ where n is the number of time units t during data gathering.

3.3.2. Semi-Markov process

The mobility behavior can be modeled as a semi-Markov process and can be applied for predicting the transition that an arbitrary user makes from its current location within time period d_t. The model assumes the knowledge of the transition probabilities; these probabilities are computed using the mobility history that is collected by each user. To avoid an opportunistic location as a FVL, we derive the frequency function *f()* that is defined as follows:

$$f(l) = \frac{n_l}{n_d} \tag{3.1}$$

where n_l and n_d denote the number of times location l is recorded in UFVLT database and the number of recorded days in UFVLT database, respectively. Thus, location l is considered FVL only and only if $f(l)$ exceeds a predefined threshold f_{th}.

Each mobile user records his mobility history (i.e., UMT and UFVLT databases); this allows for the computation of road segment (resp. FVL) transition probabilities $P_{i\to j}$ (i.e., transition from road segment (resp. FVL) i to road segment (resp. FVL) j). The prediction accuracy of road segment (resp. FVL) transitions can be improved by additionally considering prior road segment transitions of the user before the transition into the current road segment. In this case, road segment (resp. FVL) transition probability can be modified to $P_{h,i\to j}$, which is a second-order Markov Chain, where h is the sub-sequence transitions of road segments from trip origin to road segment i (resp. from FVL i to current road segment). The accuracy of the prediction can also be improved by additionally considering the type of the day and the time of the day. For example, a user who works from Tuesday to Thursday at a factory, from Friday to Saturday at a school and does not work on Monday will have three types of days: factory day, school day and rest day.

Second-order Markov Chain is derived from a semi-Markov process where the successive state occupancies are governed by the transition probabilities $P_{i\to j}$; semi-Markov process depends on both the current state and the next state transition. The semi-Markov process for a time homogeneous process is given by $Q_{i,j}(d_t)$:

$$P_{i\to j} = Q_{i,j}(d_t) = \Pr\{X_{n+1} = j, T_{n+1} - T_n \le d_t \mid X_n = i\} \tag{3.2}$$

where X_n and X_{n+1} represent the state of the system after the n^{th} and $(n + 1)^{th}$ transitions, respectively, with T_n and T_{n+1} being the times at which the n^{th} and $(n + 1)^{th}$ transitions occur, respectively. $Q_{i,j}(d_t)$ denotes the probability that, immediately after making the transition into state i, the process makes a transition into state j within t units of time. Thus, the probability $Q_{i,j}()$ (see Equation 3.2) can be computed to evaluate the predictions of an arbitrary user making a transition to a next location (e.g., FVL). However, the semi-Markov process for mobility prediction can also be extended to the case where times-of-day *TioD* (e.g., morning, noon, afternoon and night), types-of-day *TyoD* (e.g., weekend, labor day, holiday and vacation day) and the user's previous locations h are considered in the mobility pattern, i.e., extending $Q_{i,j}(d_t)$ to $Q_{h,i,j}^{TioD,TyoD}(d_t, d)$, which is defined as follows:

$$
\begin{aligned}
&P_{h,i\to j}^{TioD,TyoD} = Q_{h,i,j}^{TioD,TyoD}(d_t, d) = \\
&\Pr\left\{\begin{matrix} X_{n+1} = j, T_{n+1} - T_n \le d_t \mid X_n = i, X_{n-1} = h, \\ T_{n+1} \in TioD, T_n \in TioD, d \in TyoD \end{matrix}\right\} \\
&= \frac{n_{h,i,j}(TioD, TyoD)}{\sum_{k=1}^{N} n_{h,i,k}(TioD, TyoD)}
\end{aligned}
\tag{3.3}
$$

where d is the current date which is used to define *TyoD* at which the n^{th} and $(n + 1)^{th}$ transitions occur, $n_{h,i,j}(TioD,TyoD)$ is the number of times the transition from FVL i to FVL j has occurred within times-of-day *TioD* and types-of-day *TyoD* after crossing h; and N is the cardinality of the set of possible states j.

For the road segment transition prediction, the accuracy can be improved by additionally considering the next destination (i.e., FVL to be reached after the current trip) of the user; in this case, the probability $Q_{i,j}()$ defined in Equation (3.2) can be modified to $P_{h,i,j,D,}$ where D is the estimated destination (a FVL) of the trip and h is the set of transited road segments before entering road segment i. Thus, the semi-Markov process can also be extended to the case where the user's estimated destination is considered in the mobility prediction, i.e., extending $Q_{i,j}(d_t)$ to $Q_{h,i,j,D}^{TioD,TyoD}(d_t, d)$, which is defined as follows:

$$P_{h,i\to j,D}^{TioD,TyoD} = Q_{h,i,j,D}^{TioD,TyoD}(d_t, d) =$$
$$\Pr\left\{\begin{array}{l} X_{n+1} = j, T_{n+1} - T_n \le d_t \mid X_n = i, X_{n-1} = h, X_{last} = D \\ T_{n+1} \in TioD, T_n \in TioD, d \in TyoD \end{array}\right\} \quad (3.4)$$
$$= \frac{n_{h,i,j,D}(TioD, TyoD)}{\sum_{k=1}^{N} n_{h,i,k,D}(TioD, TyoD)}$$

where X_{last} represents the destination of the current trip, $n_{h,i,j,D}(TioD,TyoD)$ is the number of times the transition from road segment *i* to road segment *j* towards FVL *D* has occurred within times-of-day *TioD* and types-of-day *TyoD* after crossing *h*; and N is the cardinality of the set of adjacent road segments of road segment *i*. In the next subsection, we describe how our proposed model makes use of NM, UFVLT and UC databases to predict the user's destination.

3.3.3. Destination prediction model

The proposed destination prediction model (DPM), as part of DAMP, makes use of UC, NM and UFVLT databases to predict the user's destination (i.e., the next FVL to be visited). Figure 6 shows the interactions between UC, NM and UFVLT databases and DPM functions (e.g., *f(), o(), clustering , $P_{h,i,j}$(), Bel(), b() and ws()*). Indeed, the locations stored in UC and UFVLT databases represent user's potential destinations (i.e., FVLs). DPM does not predict a single destination but a cluster of destinations that likely includes the user's destination; we define a cluster of destinations as a set of FVLs that can be visited/reached by a user using the same portion of path (set of road segments) within time period d_t. Before the clustering process, we make use of the deviation function *o* to select potential destinations Ψ; the deviation function *o* measures the deviation rate of an angle (always smaller than 180) and is defined as follows:

$$o : [0,180] \to [0,1]$$
$$o(\theta) = 1 - \frac{\theta}{180}$$

Indeed, we measure the deviation rate $o(\theta_l^c)$ of each FVL *l* which is selected in UFVLT database where $f(l) \ge f_{th}$ using Equation (3.1). θ_l^c denotes the angle formed by the current movement direction (i.e., vector from trip origin to current location) and the movement

direction towards FVL *l* (i.e., vector from current location to FVL *l*). For better understanding, let us consider the example shown in Figure 7; the angle formed by the current movement direction (vector SC) and the movement direction towards FVL *L4* is the angle θ_{L4}^{c}. Notice that the maximum value of θ_{l}^{c} is 180^{o} (case of U turn).

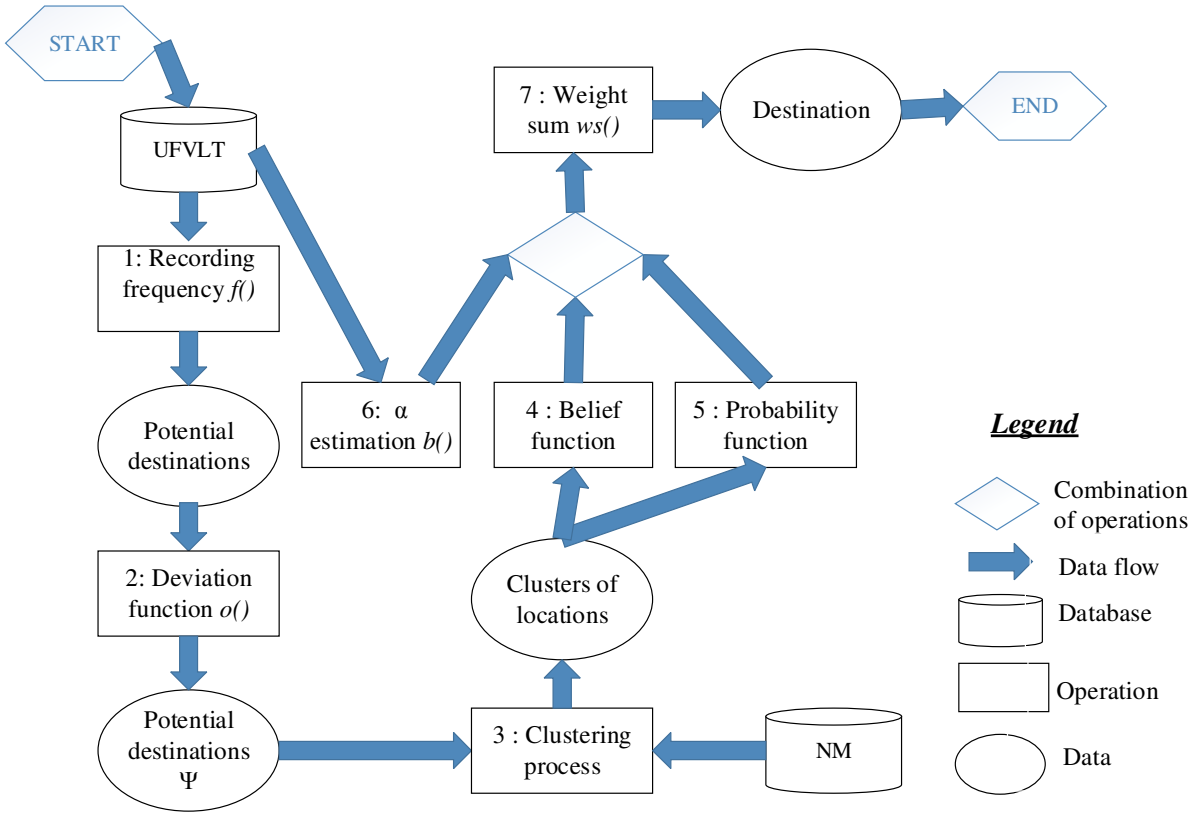


Figure 6: Illustration of DPM processes.

We define Ψ as follows:

$$\Psi = \bigcup_{l} \left\{ l \,\middle|\, o(\theta_{l}^{c}) \geq \varepsilon \right\} \tag{3.5}$$

where $\varepsilon = o(\theta_{th1})$ and θ_{th1} denotes a predefined threshold. In our proposed model, θ_{th1} is set to 90^{o} in order to consider FVLs located in front of the current location according to the current movement direction. Using the example shown in Figure 7, we compute $\Psi = \{L1, L2, L3, L4\}$ using Equation (3.5); in this case, $\varepsilon = 1/2$ assuming $\theta_{\text{th1}} = 90$. L5 is not an element of Ψ because $o(\theta_{L5}^{c}) < \varepsilon = 1/2$.

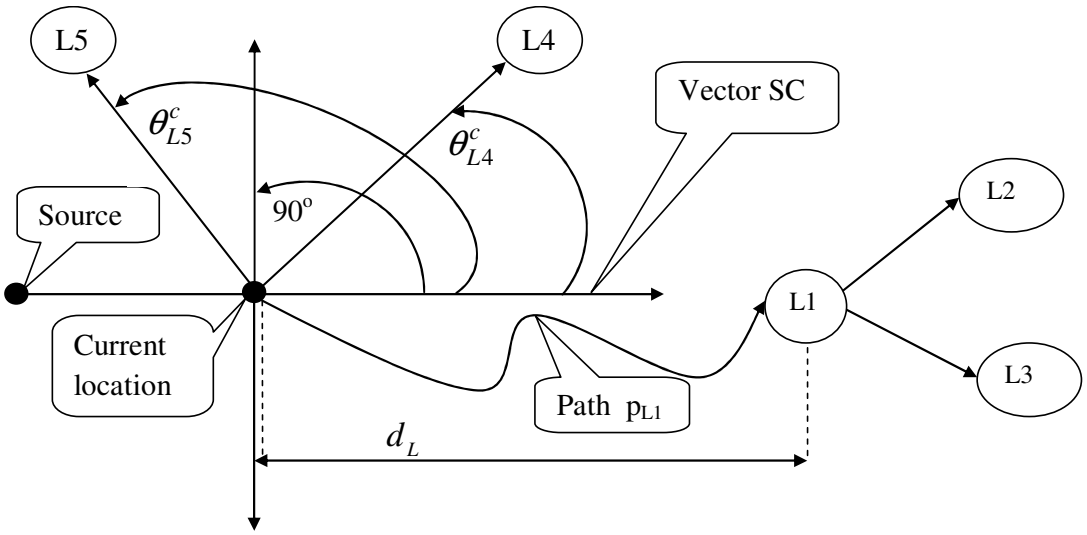


Figure 7: A destination clustering example.

After computing Ψ, we execute the clustering process. Indeed, all the elements of Ψ that may be reached using the same portion of path within a predefined time period d_t of travel form a single cluster; based on NM database, we derive a directed graph G whose edges correspond to road segments and whose vertices correspond to nodes; the road segment length represents the weight of the corresponding edge. Making use of graph G, we determine the shortest path to reach each element of Ψ using Dijkstra's algorithm; then, making use of the length of each computed path and the maximum permitted velocity of road segments, we determine the portion of that computed path after d_t time of travel; finally, the elements of Ψ, which have the same portion of the path form a single cluster. For better understanding, let us consider the example shown in Figure 7. In this example, L1, L2 and L3 may be reached using the same portion of path P_{L1} within the travel time d_L. Hence, L1, L2 and L3 form a single cluster. Intuitively, the prediction becomes more accurate when the size of clusters decreases.

DPM predicts the user's destination (or rather the FVL cluster) using a combination of belief and probability functions. DPM uses the belief function, *Bel()*, adopted in [29]. *Bel()* is based on the utilization of the mathematical theory of evidence as a tool of reasoning to investigate the user's behavior concerning his decisions about his future location (i.e., FVL). The theory is based on two ideas: the idea of obtaining degrees of belief for a related hypothesis and the idea of applying Dempster's rule for combining such degrees when they are based on different bodies of evidence; more details about this theory can be found in [29].

The net effect of Dempster's rule is that concordant bodies of evidence reinforce each other, while conflicting bodies of evidence erode each other. The main advantage of the underlying theory of evidence over other approaches [68, 69, 87] is its ability to model the narrowing of a hypothesis with the accumulation of evidence and to explicitly represent uncertainty in the form of ignorance or reservation of judgment. *Bel()* makes use of contextual information, stored in UC database, to compute the belief level, $Bel(C_i)$, of each formed cluster C_i to be the destination cluster. DPM also computes the probability $P_{h,cl\rightarrow Ci}^{TioD,TyoD}$ that the formed cluster C_i is the destination using Equation (3.3) where *cl* is the current location; it is worth noting that the set of type-of-day (i.e., *TyoD*) is different from a user to another. DPM uses a weighted sum of the belief and probability functions to compute the destination; the sum is defined as follows:

$$ws(c_i) = \alpha Bel(C_i) + (1-\alpha) P_{h,cl\rightarrow Ci}^{TioD,TyoD} \tag{3.6}$$

where α is computed as follows:

$$b : [0, +\infty] \rightarrow [0,1]$$
$$\alpha = b(n) = \begin{cases} 1-\frac{n}{RL} & if \quad 0 \le n \le RL \\ 0 & if \quad n \ge RL \end{cases} \tag{3.7}$$

where *n* denotes the number of days used to learn the user's habits and RL is the user's record lifetime. Equation (3.6) shows that as the number of learning days (of the users habits) increases, the influence of the belief function *Bel()* decreases while the influence of the probability function $P_{h,cl,Ci}^{TioD,TyoD}$ increases. DPM selects the cluster with the largest value of *ws()* as the destination cluster. DPM, which is a semi-Markov process, calculates *m* state transition probabilities; each state refers to a discrete FVL, thus, adding *O(m)* space and time complexity.

3.3.4. Path prediction model

PPM assumes a priori knowledge of the destination thanks to DPM. The operation of PPM consists of choosing a road segment (among one or more road segments) at each road junction towards the destination. The selection process starts from the current location (i.e., the road junction, immediately after the road segment where the prediction starts) and is

repeated within a predefined time period d_t of travel or until the destination (a FVL) is reached; at each occurrence of selection process, the previous road segment, that has been selected, becomes the current road segment and the road junction immediately after that current road segment becomes the current road junction. Indeed, the process terminates when a list of road segments that constitutes a path from current location to destination cluster (i.e., cluster of FVLs that may be reached using the same portion of path within a predefined time period d_t of travel) is computed. Figure 8 shows the PPM operation.

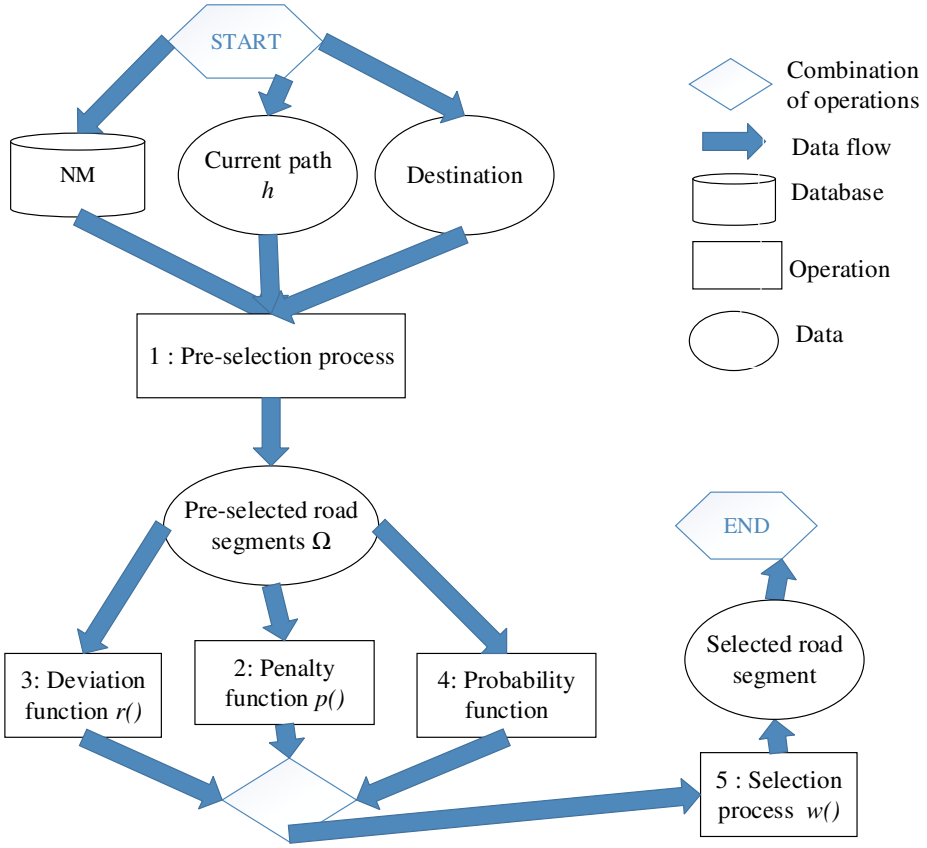


Figure 8: Illustration of PPM processes.

At each road junction (e.g., the current road junction), PPM starts by a pre-selection process choosing a set of road segments Ω among the adjacent road segments to the current road junction; Ω represents the set of potential next road segments to be visited; the pre-selection process aims to reduce the size of the set of adjacent road segments used for the selection process; it is performed making use of a deviation function r which measures the deviation rate of an angle (always smaller than 180); $r()$ is defined as follows:

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

$$r(\theta)=\begin{cases}1-\dfrac{\theta}{\Theta} & if \quad 0\le\theta\le\Theta\le 180\\ 0 & if \quad \Theta\le\theta\le 180\end{cases}$$

where Θ is angle formed by $\vec{A}$ and $\vec{B}$; $\vec{A}$ is the corresponding vector of road segment in opposite direction to the previous road segment and $\vec{B}$ is the vector from current junction to destination cluster. Then, making use of the deviation function *r()*, we measure the deviation rate $r(\theta_j)$ of each adjacent road segment *j* to the current road junction; θ_j denotes the angle formed by the corresponding vector of adjacent road segment *j* and $\vec{B}$. The pre-selected adjacent road segments *j* are those that belong to the following set:

$$\Omega=\bigcup_j\left\{j\,\middle|\,r(\theta_j)\ge\varphi\right\} \tag{3.8}$$

where $\varphi=r(\theta_{th2})$ and θ_{th2} denotes a predefined threshold. For the sake of better understanding, let us consider the example shown in Figure 9. Let $i\rightarrow C$ be the previous selected road segment and vector $\overrightarrow{iC}$ be the corresponding vector, C be the current junction, D be the destination cluster and $C\rightarrow j$ be an element of the set of adjacent segments to the current junction C. Let $\theta_{\overrightarrow{Cj}}^{\overrightarrow{CD}}$ be the angle formed by vector $\overrightarrow{CD}$ and vector $\overrightarrow{Cj}$; notice that $\overrightarrow{Ci}$ is the vector representing the road segment $C\rightarrow i$ which is in opposite direction to the previous road segment $i\rightarrow C$ that has been selected by PPM; selecting the road segment $C\rightarrow i$ as the next road segment represents a U-turn.

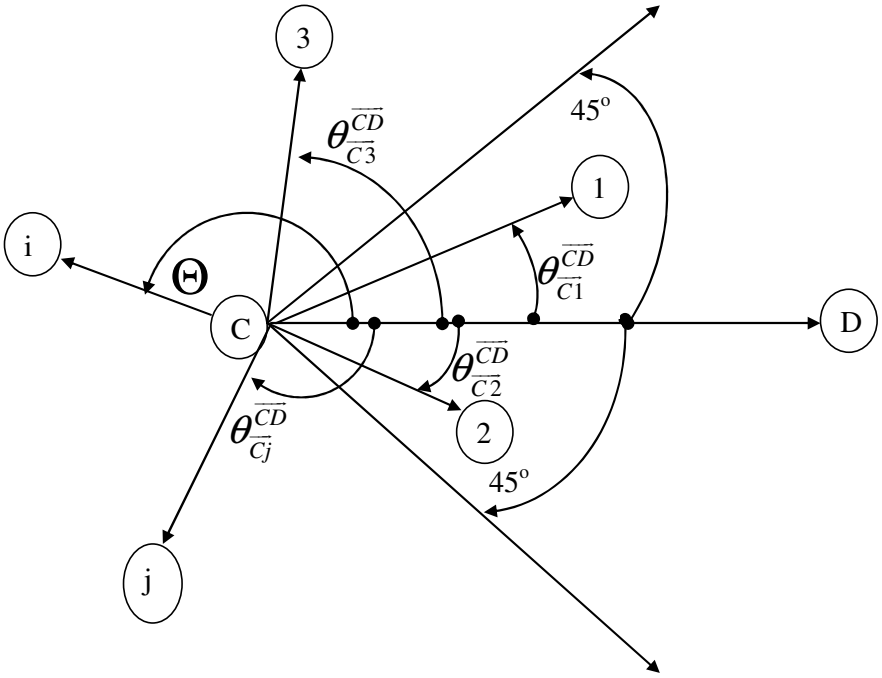


Figure 9: An example of pre-selection process.

Using the example shown in Figure 9, we compute Ω={1,2} using Equation (3.8); in this case, $\varphi = r(45)$ assuming $\theta_{\text{th2}} = 45$. The selection of a road segment from Ω as a next road segment is performed using the product of the following: (1) the transition probability $P_{h,C\to j,D}^{TioD,TyoD}$ (see Equation 3.4); (2) the deviation function *r();* and (3) the penalty function *p()* which returns one when the considered road segment may be used to reach destination cluster; otherwise, it returns null; the penalty function is defined as follows:

$$p(C \to j) = \begin{cases} 1 & if \quad possible \quad to \quad reach \quad D \quad \text{through} \quad C \to j \\ 0 & if \quad non-possible \quad to \quad reach \quad D \quad \text{through} \quad C \to j \end{cases} \tag{3.9}$$

Indeed, we compute the product function $w(C \to j)$ of each road segment $C \to j$, in Ω, using Equation (3.10) and choose the road segment with the largest value of $w(C \to j)$ as the next road segment. The product function of $C \to j$, in Ω, is defined as follows:

$$w(C \to j) = P_{ct,C\to j,D}^{TioD,TyoD} \times r(\theta_{\overrightarrow{Cj}}^{\overrightarrow{CD}}) \times p(C \to j) \tag{3.10}$$

We make use of the deviation function *r()* when computing the product function *w()* to give priority to road segments whose directions are more oriented towards the destination

cluster D (i.e., the rationale behind using angle $\theta_{\overrightarrow{Ci}}^{\overrightarrow{CD}}$ to define the deviation function). The penalty function *p()* is used to assign 0 to the product function *w()* when the considered road segment is not an option (e.g., dead end road) to reach destination cluster D. In case of lack of historical data, it will not be possible to calculate the transition probability $P_{h,C\to j,D}^{TioD,TyoD}$ (see Equation 3.4); Equation (3.10) cannot be applied to compute the product function *w()*. Thus, the selection of a road segment from Ω as a next road segment is performed using the product of the deviation function *r()* and the penalty function *p()*. In this case, the product function of $C \to j$, in Ω, is defined as follows:

$$w(C \to j) = r(\theta_{\overrightarrow{Cj}}^{\overrightarrow{CD}}) \times p(C \to j) \tag{3.11}$$

The selected road segment is added to the list of previous selected road segments; this list constitutes the predicted path from current location to destination cluster. PPM, which is a semi-Markov process, calculates g state transition probabilities. Each state refers to a discrete road segment, thus, adding $O(g)$ space and time complexity at each road junction and $O(g^2)$ for the operations of PPM.

3.4. Performance evaluation

In this Section, we evaluate, via simulations, the performance of DAMP. Similar to [20, 29, 47, 49], we define one evaluation parameter that is the prediction accuracy (i.e., path similarity). As comparison terms, we use the schemes described in [50], [47] and [20], referred to as AP1, AP2 and AP3, respectively. AP1, AP2 and AP3 were selected because, to the best of our knowledge, they represent the most recent work related to mobility prediction in MNs that outperform existing approaches (e.g., [16, 25, 49, 70]). Table 6 shows the characteristics of DAMP, AP1, AP2 and AP3.

Table 6: Prediction schemes for comparison.

Schemes	Trajectory from origin to current location	Time-of-day	Day-of-week	Only history traces	Current direction	Direction to destination
DAMP	yes	yes	yes	no	yes	yes
AP1[50]	yes	yes	no	yes	no	no
AP2[47]	yes	no	no	yes	yes	no
AP3[20]	no	no	no	yes	yes	no

3.4.1. Simulation setup

To evaluate DAMP, 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 [124]. This GPS trajectory dataset was collected in a period of over three years (from April 2007 to August 2012). 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. These trajectories were recorded by different GPS loggers and GPS-phones, and have a variety of sampling rates; 91% of the trajectories are logged every 1~5 seconds or every 5~10 meters per point. 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. According to GPS trajectories, we identify three groups: (1) subjects whose mobility is unpredictable; (2) subjects whose mobility is moderately predictable; and (3) subjects whose mobility is highly predictable. Converting GPS coordinates to Cartesian coordinates, we identify the roads by displaying all the Cartesian coordinates in a map. Algorithm 1 is used to identify the FVLs; based on the FVLs and the sequences of time-stamped points, we extract the User Contextual (UC) information. Table 7

shows the values of the parameters used in our simulations; these parameters are selected according to the road topology of the prediction area (i.e., navigation zone). For example, the parameters in Table 7 are more appropriated for a Manhattan model (i.e., a two-dimensional environment with the roads arranged in a mesh shape);

Table 7: Simulation parameters.

Parameter	**value**	**Parameter**	**value**	**Parameter**	**value**
f_{th}	1/30	θ_{th1}	90o	θ_{th2}	45o

We define one parameter, denoted by A_p, to evaluate the performance of DAMP in terms of path similarity. In the literature [20], the distance error is used to measure error for predictive path queries. However, in some cases, the distance error is small while the predicted path is very different from the actual path; this justifies taking into account path similarity to measure the performance of DAMP. Let L_{act} be the actual location of the user after travel time d_t (which will become known only in the future), L_{pred} be the location of the user in the predicted path after travel time d_t (returned by path prediction model), E_{act} be the set of road segments that the actual path from path prediction origin to L_{act} contains, and E_{pred} be the set of road segments that the predicted path from path prediction origin to L_{pred} contains. Similarly to [20], we measure path similarity A_p which is defined as follows:

$$A_p(E_{act}, E_{pred}) = \frac{2.\left|E_{act} \cap E_{pred}\right|}{\left|E_{act}\right| + \left|E_{pred}\right|} \tag{3.12}$$

In the remainder of this paper, the terms path similarity and accuracy will be interchangeably used. Unless stated otherwise, in all simulation scenarios, we use the two months of the Microsoft Research Asia laboratory's dataset (June-July 2012) to learn users' habits (in this case we state that the length of the learning phase is 60 days); the learning phase denotes the period between the time at which the mobility data collection is started and the time at which the prediction is performed; the prediction phase comes after this phase; we use the last month of this dataset (August 2012) as prediction phase; this period is also used to compare the actual and predicted paths similarity. We also assume that the path from trip

origin to the current location (where the prediction process is executed) corresponds to 500 meters of the path from trip origin to destination and the prediction length d_t is 2 hours.

In order to compare DPM with the approach proposed in [29], we compute the accuracy of destination prediction *Ad*, which is defined as follows:

$$A_d = \frac{n_{bp}}{n_{tp}} \tag{3.13}$$

where n_{bp} and n_{tp} denote the number of correct estimates (i.e., L_{act} and estimated destination are the same) and the total number of estimates, respectively.

3.4.2. Results analysis

Simulation results are averaged over multiple 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 the path towards this destination. For each run, we compute *Ap* (resp. *Ad*) using Equation (3.12) (resp. 3.13); thus, to obtain the simulation results shown in Figure 10, 11, 12 and 13 we compute the average of the *five hundred* runs.

Figure 10 shows the average accuracy of destination prediction when varying the length of the learning phase. We observe that for DAMP (resp. DPM-Samaan [29]), the average accuracy of destination prediction increases (resp. remains constant) with the length of the learning phase. This can be explained by the fact that DAMP uses historical data to perform destination prediction; indeed, the longer the length of the learning phase, the better the knowledge about users' mobility habits, and ultimately the higher the prediction accuracy; in contrast, DPM-Samaan uses only the user context (his goals and interests). Figure 10 also shows that DAMP outperforms DPM-Samaan; DAMP provides an average accuracy of 0.94 per 5 days of learning phase while DPM-Samaan provides an average of 0.75 per 5 days of learning phase; overall, the average relative improvement (defined as [average *Ad* of DAMP - average *Ad* of DPM-Samaan]) of DAMP compared to DPM-Samaan is about 19%. At 0 day of learning phase length (i.e., in case of lack of historical data), DAMP uses only the belief function of DPM-Samaan according to Equations (3.6) and (3.7); yet, DAMP and DPM-

Samman do not provide the same accuracy; this can be explained by the fact that DAMP performs the destination selection process after pre-selection and clustering processes. At 30 days of learning phase length, DAMP accuracy is about 96%; according to Equations (3.6) and (3.7), at 30 days of learning phase length, DAMP does not make use of the belief function proposed in [29]; indeed, for *RL*=30 days (the maximum value of RL) and *n*=30 days (which represents 30 days of learning phase), Equation (3.6) becomes $ws(c_i) = P_{h,cl \to Ci}^{TioD,TyoD}$; in this case, DAMP uses only the probability function.

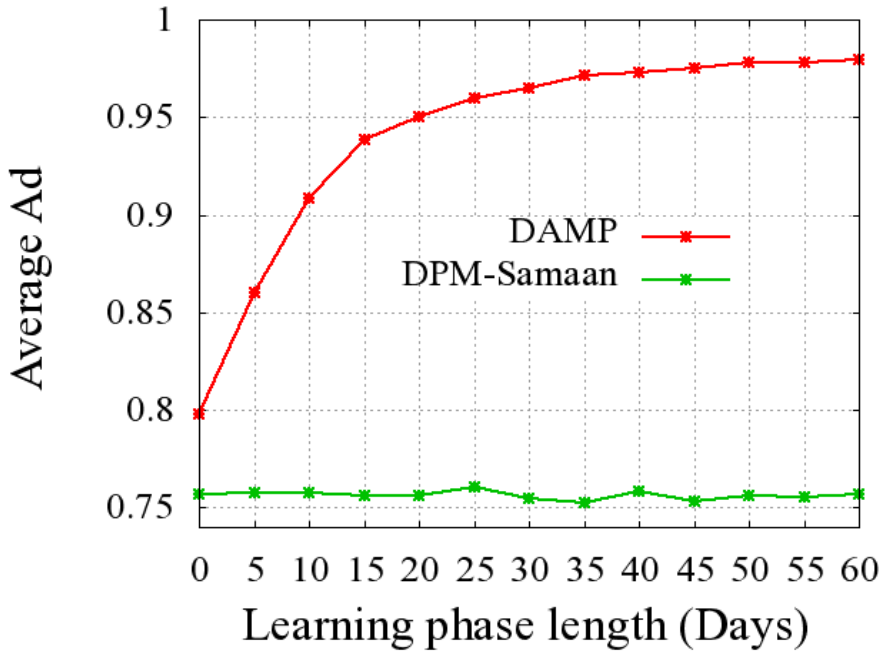


Figure 10: Average destination prediction accuracy versus learning phase length variation

Figure 11 shows the average accuracy when varying the length of the learning phase. We observe that, for the four schemes, the average accuracy increases with the length of the learning phase. This is expected since the longer the length of the learning phase, the better the knowledge about users' mobility habits, and ultimately the higher the prediction accuracy becomes when historical data is used to perform prediction. Figure 11 also shows that DAMP outperforms AP1, AP2 and AP3; for example, DAMP provides an average accuracy of 0.84 per 5 days of learning phase length while AP1 (more efficient than AP2 and AP3 in this scenario) provides an average of 0.54 per 5 days of learning phase length; overall, the average relative improvement (defined as [average *Ap* of DAMP - average *Ap* of AP1]) of DAMP

compared to AP1 is about 30%. At 10 days of learning phase length, DAMP accuracy is about 72% while AP1 requires 60 days of learning phase length to provide the same accuracy; this means that DAMP requires a smaller learning phase length (1 day compared to 6 days for AP1) to perform prediction with similar accuracy. This can be explained by the fact that (1) DAMP filters data taking into account the type of day (e.g., Labor Day, Weekend and Weekday) and (2) DAMP makes use of user context in addition to the mobility history traces. Indeed, user context along with historical data helps improving prediction accuracy. We also observe that AP1 outperforms AP2 and AP3; this can be explained by the fact that AP1, for prediction purposes, uses data filter (i.e., time-of-day); thus, when the length of the learning phase increases, the prediction accuracy increases; in this case, AP1 becomes more accurate than AP2 and AP3 when the length of the learning phase exceeds 5 days. In case of lack of historical data (i.e., learning phase length is equal to 0 day), DAMP accuracy is about 38% while AP1, AP2 and AP3 accuracy is 0%; this can be explained by the fact that DAMP makes use of direction towards destination when there is no mobility history traces (see Equation 3.11); AP1, AP2 and AP3 are only based on historical data; so, without mobility history traces, they cannot perform a prediction.

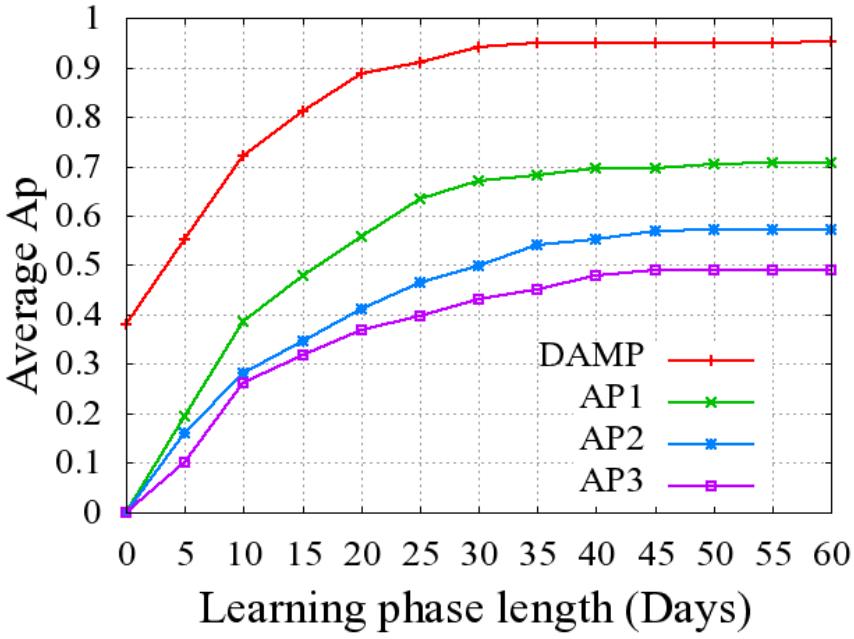


Figure 11: Average prediction accuracy versus learning phase length variation.

Figure 12 shows the average accuracy when varying the length of the path from trip origin to current location. We observe that for DAMP, AP1 and AP2 (resp. AP3), the average accuracy increases (resp. remains constant) with the length of the path from trip origin to current location (i.e., path already traveled by the user towards destination); this can be explained by the fact that DAMP, AP1 and AP2 use the portion of the path already traveled by the user to compute the remaining path to destination; in contrast, AP3 uses only the last crossed location; indeed, DAMP, AP1 and AP2 try to match the current path (from trip origin to current location) to paths stored in their databases of user's trajectory history. Thus, when the size of the current path increases, the prediction accuracy increases. Figure 12 also shows that DAMP outperforms AP1, AP2 and AP3; DAMP provides an average accuracy of 0.97 per 0.25 kilometer of path already traveled while AP1 (more efficient than AP2 and AP3 in this scenario) provides an average of 0.83 per 0.25 kilometer of path already traveled; the average relative improvement of DAMP compared to AP1 is about 14%. We also observe that DAMP's (resp. AP1's) average accuracy increases more rapidly between 0 and 0.5 kilometer (resp. between 0.5 and 1.5 kilometers) of path already traveled by the user. This means that DAMP requires a smaller path already traveled by the user (about 0.5 kilometer compared to 1.5 kilometers for AP1) to predict path with better accuracy. We also observe that, at 0 kilometer of path already traveled (i.e., the prediction process is executed at trip origin), DAMP performance is about 96% while AP1 requires 1.25 kilometers of path already traveled to provide the same performance. This means that AP1, compared to DAMP, requires that the user be located more closely to the destination to predict path with better accuracy. This can be explained by the fact that (a) DAMP uses the direction from current location to destination to compute, at each location, the potential next location; and (b) DAMP filters historical data based on the type of the day.

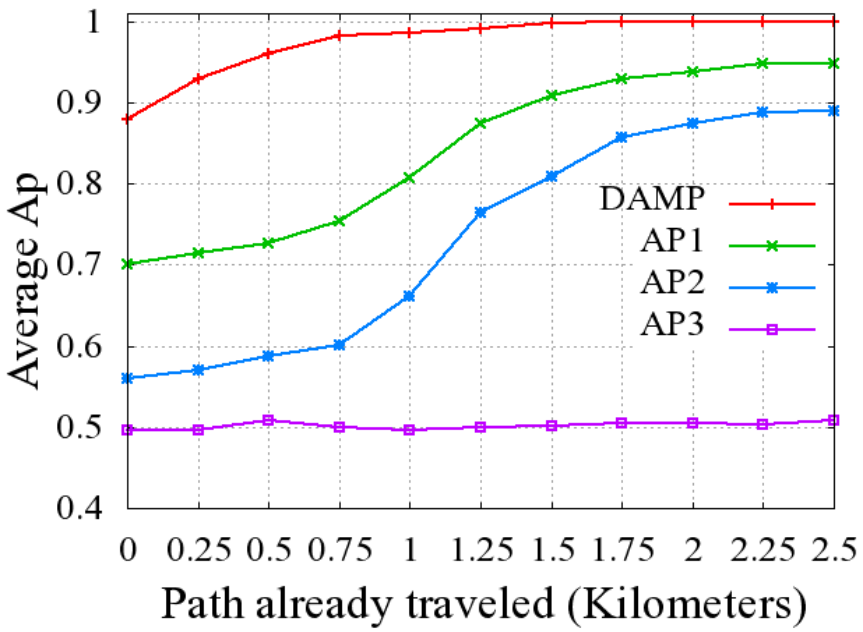


Figure 12: Average prediction accuracy versus path already traveled variation.

Figure 13 shows the average accuracy prediction when varying the prediction length (i.e., d_t). We observe that, for the four schemes, the average accuracy decreases with the length of the prediction. This is expected since when the prediction length increases, the number of possible paths increases and thus the prediction accuracy decreases. Figure 13 also shows that DAMP outperforms AP1, AP2 and AP3; this is mainly due to the fact that DAMP predicts a destination cluster (in opposition to a single destination) and makes use of the movement direction towards the destination cluster during the prediction process. In particular, destination clustering allows the grouping of probable destinations (as a single destination: a cluster) when the length of prediction increases. This grouping reduces the number of probable destinations and increases the path prediction accuracy towards these probable destinations. DAMP provides an average accuracy of 0.93 per 0.5 hour of prediction length while AP1 (more efficient than AP2 and AP3 in this scenario) provides an average of 0.60 per 0.5 hour of prediction length; overall, the average relative improvement of DAMP compared to AP1 is about 33%. We also observe that AP2 (resp. AP3) outperforms AP1 around 3.5 (resp. 4) hours of prediction length. This can be explained by the fact that AP2 and AP3, in contrast to AP1, for prediction purposes, consider user's current direction (i.e., vector/direction from trip origin to current location). Thus, when the prediction length increases, the prediction accuracy of

AP2 and AP3 are less impacted compared to AP1; indeed, the usage of current direction reduces the number of probable paths and increases the path prediction accuracy; in this case, AP2 (resp. AP3) becomes more accurate than AP1 when the length of the prediction exceeds 3.5 (resp. 4) hours. Even though AP2 and AP3 use similar techniques as DAMP (e.g., user's current direction), they do not make use of direction towards the destination. Indeed, when the length of the prediction exceeds 3.5 (resp. 4) hours, DAMP provides an average of 0.88 per 0.5 hour of prediction length while AP2 (resp. AP3) provides an average of 0.49 (res. 0.44) per 0.5 hour of prediction length; overall, the average relative improvement of DAMP compared to AP2 (resp. AP3) is about 39% (resp. 44%). This means that DAMP predicts remaining path to destination with better accuracy despite the expansion of length of prediction. This can be explained by the fact that DAMP uses the direction from current location to destination cluster.

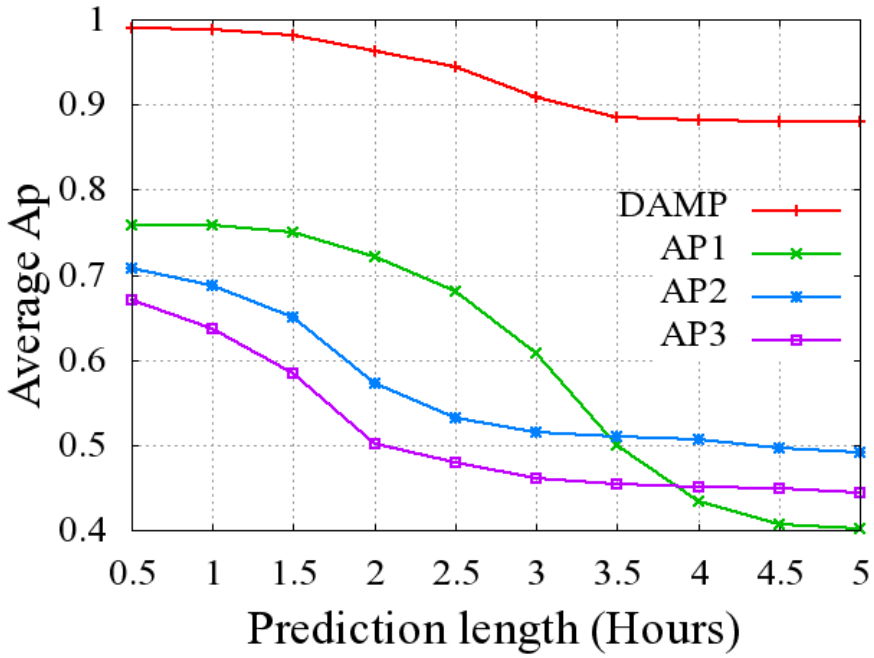


Figure 13: Average prediction accuracy versus prediction length variation.

In summary, the analysis of the simulation results shows that schemes which use data (e.g., user context) in addition to historical mobility traces outperform schemes which are limited to mobility traces when the size of historical mobility traces is not large enough (i.e., the length of the learning phase is not long enough). Likewise, historical mobility traces filtering increases the accuracy when the length of the learning phase increases. We also

observe that schemes which consider the path from trip origin to current location outperform others when the length of path already traveled increases while the schemes which consider the current movement direction (i.e., vector/direction from trip origin to current location) outperform others when the length of prediction increases. Finally, taking into account the direction from current location to destination allows for improving performance despite the expansion of prediction length. We summarize DAMP evaluation findings as follows: (1) DAMP uses path from trip origin to current location in the prediction process; even though AP1 and AP2 use path from trip origin to current location, they require a long path from trip origin to current location to predict, with better accuracy, the path from current location to destination; (2) DAMP uses movement direction in the prediction process; even though AP2 and AP3 use movement direction, they do not consider the direction towards destination; (3) DAMP uses user context and filters historical data based on the type of the day and the time of the day; this helps increasing accuracy. Even though AP1 uses similar data filter, it is limited to the days-of-week, and thus, it requires a long time (4 times more than DAMP) to predict, with better accuracy, the path from current location to destination.

The communication complexity of DAMP is 0; indeed, DAMP does not communicate with the network system to predict users' mobility; all the databases and the processes are maintained/run by the user equipment UE. The computational complexity of DAMP is $O(m)+O(g^2)$ where m is the number of FVLs and g is the number of road segments of the navigation map (NM). Indeed, DPM calculates m transition probabilities of FVL while PPM calculates g^2 transition probabilities of road segment. We evaluate the computational complexity as a computation time on a Samsung Galaxy S4 (2GB of RAM, 4*1.9 GHz of processor speed); The results show that the computation time is less than 1 second for 54% of users, within 1 second and 2 seconds for 31% of users, and within 2 seconds and 3 seconds for 15% of users.

3.5. Conclusion

In this paper, we introduced a destination and mobility path prediction model, called DAMP, for predicting subsequent transitions of road segments across the mobility of users within a predefined time period d_t. DAMP consists of two models: DPM (for predicting user's

destination) and PPM (for predicting subsequent transitions of road segments towards predicted destination). We evaluated, via simulations, DAMP and compared it against three related schemes recently proposed in [20, 47, 50]. The simulation results demonstrated that DAMP achieved better accuracy regardless of the predictability level of users, learning phase length, prediction lengths, and already traversed path length. The obtained results also clearly show that the utilization of user context, path traversed from trip origin to current location and movement direction together with fine-grain filtering of historical data (e.g., type of day) greatly increases path prediction accuracy. The findings of this contribution (estimated path) can be used, for example, to better estimate the handoff times along estimated paths [31]. Currently, we are working on integrating the proposed DAMP with a suitable bandwidth-management and admission control scheme; a preliminary version of this scheme can be found in [107].

Chapitre 4 :

Mobility Prediction-aware Bandwidth Reservation Scheme for Mobile Networks

Apollinaire Nadembéga, Abdelhakim Hafid, Tarik Taleb

Abstract

Bandwidth is an extremely valuable and scarce resource in mobile networks; therefore, efficient mobility-aware bandwidth reservation is necessary in order 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 Mobility Prediction aware Bandwidth-Reservation scheme (MPBR). 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 scheme, called HTE, that aims to estimate the time windows when a user will perform handoffs along the path to his destination; (2) an available bandwidth estimation scheme, called ABE, 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 scheme, called ECaC, that aims to control bandwidth allocation in the network cells. The simulation results show that MPBR outperforms existing schemes [28, 97, 98] in terms of reducing handoff call dropping rate.

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

Status: This article is submitted to IEEE Transactions on Vehicular Technology 2013; it is based on the following published paper:

A. Nadembéga, A. Hafid and T. Taleb. *Handoff Time Estimation Model for Vehicular Communications*, IEEE ICC, Budapest, Hungary, Jun. 2013.

4.1. Introduction

Applications, such as video streaming, IPTV, and VoIP, are increasingly prevalent over telecommunication networks; thus, it becomes important to provide the quality of service (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 WiMAX and 3GPP accesses, could offer anytime and anywhere access to mobile users [16, 30, 31, 101, 103, 125]. 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 [16]. 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 [31]. Furthermore, call admission control (CAC), at the level of each cell towards the destination, is needed to decide whether or not to accept a call into the corresponding cell [16, 101, 103, 125]. 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 hand-offing to this cell [21, 103, 104]. 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 MPBR, that provides QoS to mobile users while maintaining efficient bandwidth utilization. MPBR consists of three schemes: (1) a handoff time estimation scheme, called HTE; (2) an available bandwidth estimation scheme, called ABE; and (3) an efficient call admission control 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 [31], 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 [21, 30] can be used to predict the path for a given user). 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 distributions (PDF) of travel times between two arbitrary locations l_1 and l_2 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); 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 behaviour 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 minutes) 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). 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 knowledge of the authors, the proposed approach (MPBR) is the first to consider estimating/predicting available bandwidth in cells to be traversed by new calls of mobile users in order 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 smart phones 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., [126-128]).

The remainder of this paper is organized as follows. Section 4.2 presents related work. Section 4.3 presents a description of the handoff time estimation scheme (HTE), the available bandwidth estimation scheme (ABE) and the efficient call admission control scheme (ECaC). Section 4.4 evaluates, via simulations, the proposed MPBR. Finally, Section 4.5 concludes the paper.

4.2. Related Work

The schemes proposed in [14, 27, 94, 97, 98, 102-106] decide to accept a new call or not based on the behavior/state of the source cell and are usually simpler to implement but not efficient. On the other hand, predictive mobile-oriented schemes [27, 28, 89, 93, 95, 96, 100, 101] 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. Vassilya and Isik [16] 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 non-distributed schemes [26, 28, 86, 89, 92-97] and two or more cells for distributed schemes [27, 98]) and the way handoff requests are handled (e.g., non-prioritized or prioritized handoff). Non-prioritized handoff CAC schemes [129] 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 [26-28, 86, 89, 92, 93, 95, 97-103] give handoff calls precedence over new calls (i.e., reject a new call in order to accommodate a handoff call); many attempts were made to address the issue of prioritized handoff CAC making use of users mobility prediction (i.e., predictive mobile-oriented and prioritized handoff CAC schemes [27, 28, 89, 93, 95, 100, 101]). Thus, 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. 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 [17]. Having this knowledge in advance is not possible in realistic scenarios [16]; 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) handoff time estimation: 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) call admission control: 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 [28, 97, 98] 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, Jun*, et al.* [98] proposed a CAC and bandwidth reservation scheme which is cell-oriented, distributed and supports prioritized handoff. The current available bandwidth is estimated 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 the fact that using historical network bandwidth observations (cell behavior/state) does not provide an accurate estimation of available bandwidth compared to using individual users' behaviors. Wee-Seng, *et al.* [28] proposed a CAC and bandwidth reservation scheme which is predictive mobile-oriented, non-distributed 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 probability density function 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 source cell; otherwise, it is blocked. This scheme suffers from two key limitations, namely (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 call admission control 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. Jung-Shyr Wu*, et al.* [97] 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. 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 (a) the available bandwidth is sufficient to accommodate the call in the source cell; and (b) 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 [28], the call admission control proposed in [97] 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 [17, 98] 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 towards destination; (2) The schemes that make use of prediction techniques either require additional equipment [26, 27], generate a significant traffic overhead in terms of mobility data exchanges between users and network backbone [49], do not consider stop durations and traffic lights [17, 18, 20, 26-28, 49], make use of old road traffic data [18, 28] or rely only on historical data about previous users [28]; (3) Their admission control procedures are limited to the next cell [26-28, 86, 97]; the other cells in the path to destination are not considered; and/or (4) they rely only on historical network bandwidth observations [98].

In this paper, we propose a scheme, to process call requests, that incorporates solutions to the above-mentioned limitations. In this paper, we use path prediction schemes proposed in [21, 30]; 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 handoff time estimation; such estimation will allow to compute, with some accuracy, entry/exit times into/from cells along the path from source to destination; this will help computing available bandwidth and deciding to accept/reject calls.

4.3. Predictive Mobile-Oriented Bandwidth Reservation Scheme

In this section, we present the details of the mobility prediction-aware bandwidth reservation scheme (MPBR). More specifically, we present the details of (1) the handoff time estimation scheme (HTE) that estimates the time windows when a user will perform handoffs along his movement path to destination; (2) the available bandwidth estimation scheme (ABE) that estimates the available bandwidth in advance in a cluster of cells using the estimated handoff times; and (3) the efficient CAC scheme (ECaC) that controls, given the estimated available bandwidth computed by ABE, bandwidth allocation. Table 8 shows the list of symbols/variables that are used to describe the proposed scheme.

Table 8: Summary of notations

Symbol	Description	Symbol	Description
l, l_s	Length of road segment	Δt	Travel time on road segment
$n(t)$, $d(t)$	Number of users and density of road segment at time instant t	Δd	Stopping time at road junction
q_i, C	Time of color I and Traffic light cycle time	ΔT	Travel time on road segment with stopping time
d_1 , d_2	Lower and upper boundary density values	$F_{l1,l2}$ ()	CDF of transit/travel times of link (l_1, l_2)
a, d, v_m	Acceleration, deceleration and constant velocity	$[T_0,T_z]$, z	Estimation time interval and number of time unit within this interval
Δt_a l_a	Travel time and distance of acceleration phase	$pbw_{alloc,l}^{i,T_k}$	Amount of passive allocated bandwidth to call l at T_k
Δt_d, l_d	Travel time and distance of deceleration phase	pbw_{alloc}^{i,T_k}	Amount of passive allocated bandwidth to user u_i at T_k
Δt_c, l_c	Travel time and distance of constant phase	$PBW_{ava}^{c_j,T_k}$	Estimated available bandwidth in cell Cj
t_i^l, t_i^u	Lower and upper values of estimated handoff time of cell C_i	$\left[t_j^a, t_j^d\right]$	Time interval user u will spend in cell Cj

4.3.1. Handoff time estimation

4.3.1.1. Traffic flow and queuing models

4.3.1.1.1. 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→b ≠ b→a). We refer to the intersection of a road and the border of a cell as a handoff point.

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:

$$f(t) = d(t) \times v(t) \tag{4.1}$$

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

$$d(t) = \frac{n(t)}{l} \tag{4.2}$$

whereby $n(t)$ 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:

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 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 [58]; indeed, driving behavior is a cycle of acceleration, maintaining of a constant velocity, deceleration and finally the stopping. The free flow velocity is not the same for all users. Figure 14 illustrates a simplified driving behavior cycle and the length of road segment portion associated to each phase of the cycle. In Figure 14(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.

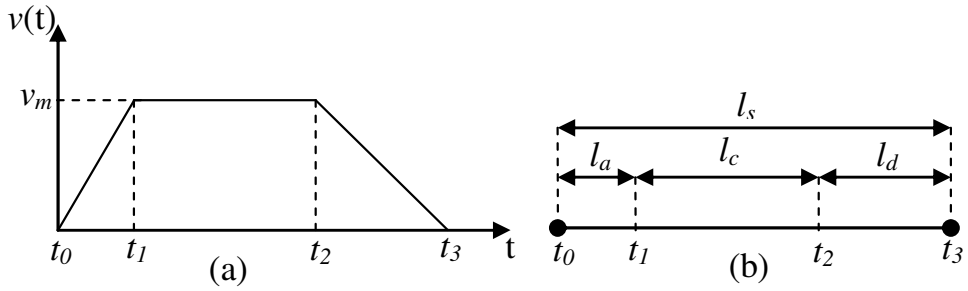


Figure 14: (a) Simplified driving behavior cycle and (b) length of road segment portion associated to each phase of the driving behavior cycle.

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 9). In case of lack of traffic lights, we apply the "first come, first serve" approach.

Table 9: 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

4.3.1.1.2. 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, under-saturated 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 (i) *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; (ii) *under-saturated 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 (iii) *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 under-saturated 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 (resp. the absence) of a queue at intersection. Indeed, in the free flowing condition, there is no queue while in the under-saturated 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:

$$v(t_i) = \sigma \times (t_i - t_{i-1}) + v(t_{i-1}) \tag{4.3}$$

whereby σ and $(t_i - t_{i-1})$ denote the acceleration/deceleration (depending on the movement) and the minimum time granularity, respectively. Based on Equation (4.3) and the simplified driving behavior cycle shown in Figure 14, we derive the expression of travel time of acceleration phase Δt_a and deceleration phase Δt_d as follows:

$$\Delta t_a = v_m / a \text{ and } \Delta t_d = v_m / d \tag{4.4}$$

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; that is given by:

$$l_c = l_s - (l_a + l_d) \tag{4.5}$$

where l_s denotes the length of the road segment. 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 Equation 4.3) and given by:

$$\begin{aligned} l_a &= \left(a/2 \times (\Delta t_a)^2\right) \\ l_d &= \left(d/2 \times (\Delta t_d)^2\right) + \left(v_m \times \Delta t_d\right) \end{aligned} \tag{4.6}$$

Using Equation (4.5), the expression of travel time of a constant velocity phase Δt_c is as follows:

$$\Delta t_c = l_c / v_m \tag{4.7}$$

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:

$$\Delta t = \Delta t_a + \Delta t_c + \Delta t_d \tag{4.8}$$

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

$$\Delta d = \sum_{\omega=1}^{i} \omega \Delta d_\omega \Big/ \sum_{\omega=1}^{i} \omega \tag{4.9}$$

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 ω^{th} 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 case of 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 9, we derive the expression of the remaining time of color *i* when the user reaches the position of traffic light as follows:

$$r = q_i - \left[\left(t_s \bmod C \right) - \sum_{f=1}^{i-1} q_f \right] \tag{4.10}$$

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

$$\Delta d = \begin{cases} 0 & \text{if non stop} \\ r & \text{if stop is required} \end{cases} \tag{4.11}$$

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

$$\Delta d = m \times \Delta D \tag{4.12}$$

where m and ΔD denote the length of the front queue and the average stopping time of the considered stopping position, respectively. Using Equations (4.8), (4.9), (4.11) and (4.12) (depending on the condition and the case), we derive the travel time on road segment with stopping time as follows:

$$\Delta T = \Delta t + \Delta d \tag{4.13}$$

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. Also, 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:

$$\Delta T = t_e - t_a \tag{4.14}$$

4.3.1.2. Databases

To implement HTE, we assume that the User Equipment (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 movement; 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): it 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 [21, 30] 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 within its radio coverage area; an entry/record in NM consists of first intersection/handoff point ID_1, second intersection/handoff point ID_2, velocity v, road segment ID and the length l of the road segment ID formed by the location pair (intersection/handoff point ID_1, intersection/handoff point ID_2) where the velocity limit is v.

(2) UE embeds GPS, stopwatch and accelerometer; indeed, at regular epoch $\Delta T'$, GPS samples the user's current location, stopwatch samples his stopping times while accelerometer samples his acceleration and his velocity, together with the timestamp.

To limit the size of DDS and DST, they are deleted when the user reaches the destination.

4.3.1.3. 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, DDS 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 Equation (4.2), 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 towards the destination.

We first propose estimating the PDF of road segment transit/travel times by mobile users. The transit time, by a user travelling on road segment S towards 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: 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 Equations (4.8), (4.9), (4.10), (4.11) and (4.13). Indeed, Equation (4.8) is derived using

Equations (4.4) - (4.7); based on last values of acceleration, deceleration and constant velocity of users stored in database DDB, we compute travel time during these three phases ,using Equation (4.4); then; using the constant velocity stored in database DDB and travel distance during the constant phase, we compute travel time during this phase using Equation (4.7); notice that travel distance during the constant phase is derived using Equation (4.5) where travel distance during the acceleration and deceleration phases are computed using Equation (4.6) and the last values of acceleration, deceleration and constant velocity of users stored in database DDB; Equations (4.4) and (4.6) derive from the physical law of movement. To compute stopping times, Equation (4.9) makes use of database DST while Equation (4.11) uses Equation (4.10) where we assume that traffic light cycles are known a priori; arrival time t_s is the median of CDF of travel times $F_{l_1,l_2}(.)$ which is defined below. The expression of t_s is as follows:

$$t_s = F^{-1}_{l_1,l_2}(0.5) \tag{4.15}$$

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

(2) Population in the under-saturated condition: 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 Equations (4.8), (4.12) and (4.13). Equation (4.8) is used in the same way as the free flow condition. For Equation (4.12), *AD* is computed based on database DST while *m* is assumed to be 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: the times to transit S by users who have already transited S and are currently located on adjacent road segments towards their destinations. Indeed, based on arrival times and exit times of users stored in database DDB, HTE computes travel times making use of Equation (4.14).

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:

$$p_S(\Delta T_i) = \frac{n_S^{\Delta T_i}}{n_S} \tag{4.16}$$

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, 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:

$$p_{l_1,l_2}(\Delta T) = \left(\prod_{i=1}^{m} p_{S_i} \right)(\Delta T) \tag{4.17}$$

where m denotes the number of road segments S_i forming the link (l_1, l_2). Using Equation (4.17), we derive the cumulative distribution function (CDF) of transit/travel times of link (l_1, l_2) $F_{l_1,l_2}(.)$ as follows:

$$F_{l_1,l_2}(\Delta T') = \sum p_{l_1,l_2}(\Delta T \le \Delta T') \tag{4.18}$$

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 (lc, h_k) (i.e., between the current location lc and the handoff point h_k) of transit/travel times by mobile users $F_{lc,h_k}(.)$. 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 (lc, 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 $F_{lc,h_k}(.)$ and given by:

$$\forall \delta_u, \delta_l \in [0;1] F_{lc,h_k}^{-1}(1-\delta_l) = \Delta t_{h_k}^l ; F_{lc,h_k}^{-1}(\delta_u) = \Delta t_{h_k}^u \quad (4.19)$$

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:

$$t_k^l = t_0 + \Delta t_{h_k}^l \text{ and } t_k^u = t_0 + \Delta t_{h_k}^u \quad (4.20)$$

where t_0 denotes the initial time of the estimation.

4.3.2. 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.

In this section, we describe the details of ABE and explain how the *n*-tuple Ω predictions, computed by HTE (see Section 4.3.1), 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.

4.3.2.1. System model

Similar to [89], we do not consider delay-insensitive calls that can tolerate long handoff delays, as well as, soft handoffs in CDMA 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} [28, 89]. 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}.

4.3.2.2. Databases

We make the following two assumptions:

(a) The network maintains a database, called User Calls 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) A prior knowledge of the time windows when a user will perform handoffs along the predicted path to a destination $\Omega = \left\langle \left(t_1^l, t_1^u, c_1\right), ..., \left(t_j^l, t_j^u, c_j\right), ..., \left(t_n^l, t_n^u, c_n\right) \right\rangle$.

4.3.2.3. Description

In order 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 $C = \{c_1, ..., c_j, ..., c_n\}$ 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 $T_k = T_0 + k\Delta t$ where $k \in N$ and Δt denotes the time unit of estimation; the index z denotes the number of time units within time interval $[T_0, T_z]$.

Let $U = \{u_1, ..., u_i, ..., u_m\}$ 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 the amount of passive allocated bandwidth pbw_{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:

$$pbw_{alloc}^{i,T_k} = \sum_{l=1}^{q} pbw_{alloc,l}^{i,T_k} \tag{4.21}$$

where q is the number of calls of u_i that are expected to be ongoing at T_k and $pbw_{alloc,l}^{i,T_k}$ is the amount of passive allocated bandwidth to call l at T_k. Using Equation (4.21), 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 10).

Table 10: Matrix of passive allocated bandwidth.

time \ users	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}		

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 $c_j \in C : Soj = \{..., (t_j^a, t_j^d, c_j), ..., (t_n^a, c_n)\}$. Using *Soj* and the matrix of passive allocated bandwidth (see Table 10), we compute the estimated available bandwidth in each cell $c_j \in C$ at T_k as follows:

$$PBW_{ava}^{c_j,T_k} = BW^{c_j} - \sum_{i=1}^{g} pbw_{alloc}^{i,T_k} \tag{4.22}$$

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 Equation (4.22), we compute the matrix of estimated available bandwidth (see Table 11).

Table 11: Matrix of estimated available bandwidth.

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

4.3.3. Call admission control

In order to better understand the logic leading to the Efficient Call Admission Control (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 in order 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.

4.3.3.1. Databases

We assume (a) that the network maintains a database, called Earlier Completed Calls Data (ECCD), which records data about earlier completed calls; an entry/record in ECCD contains call ID, bandwidth *b*, time *t1*, time *t2,* cell ID and date *d*; the call is initiated with *b* as allocated bandwidth and completes in cell ID, at *d*, at *t1* while it was estimated/predicted to complete at *t2* ($t2 > t1$). 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; (b) a prior knowledge of the time windows when the user will perform handoffs along the predicted path

to a destination $\Omega = \left\langle \left(t_1^l, t_1^u, c_1\right), ..., \left(t_j^l, t_j^u, c_j\right), ..., \left(t_n^l, t_n^u, c_n\right) \right\rangle$ (computed by HTE); and (c) a prior knowledge of the matrix of estimated available bandwidth (computed by ABE).

4.3.3.2. Description

ECaC uses ECCD, Ω and the matrix of estimated available bandwidth (see Table 11) to manage the bandwidth allocation in the cells along the path to the destination of the user in question. In order 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, $C_u = \{c_1, ..., c_j, ..., c_w\}$ the list of cells to be traversed by the user u to destination, $\Omega_u = \left\langle \left(t_1^l, t_1^u, c_1\right), ..., \left(t_j^l, t_j^u, c_j\right), ..., \left(t_w^l, t_w^u, c_w\right) \right\rangle$ the handoff time windows of the user u along the path to destination, and $Soj_u = \left\langle ..., \left(t_j^a, t_j^d, c_j\right), ... \left(t_w^a, c_w\right) \right\rangle$ 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 $PBW_{ava}^{c_j, T_k}$ (i.e., the amount of bandwidth that should not be reserved or used in cell $c_j \in C_u$ at $T_k \in \left[t_j^a, t_j^b\right]$) 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 Figure 15. 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 Figure 15). 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 user has to wait before successfully starting a new call. The expression of BIS is given by:

$$BIS = t_j^a - T_0 \tag{4.23}$$

where T_0 denotes the time of the call request and t_j^a (e.g., t_3^a in Figure 15) the time when the user is expected to enter cell c_j (e.g., cell 3 in Figure 15). 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 remainder of this section, "current user" and "user in question" (i.e., user who makes the new call request) are interchangeably used, referring to the same user.

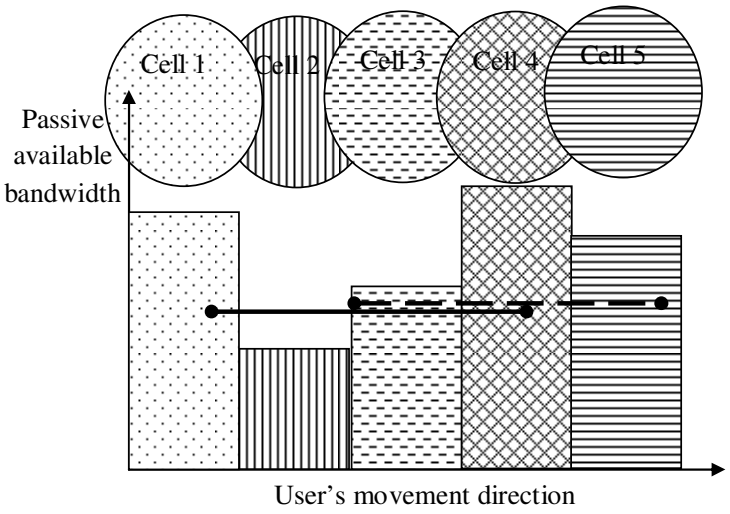


Figure 15: Illustration of a user's required bandwidth and estimated available bandwidth along the user's path to destination.

In order 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 is exceeds BISt. CPE is computed as the sum of the following measurements:

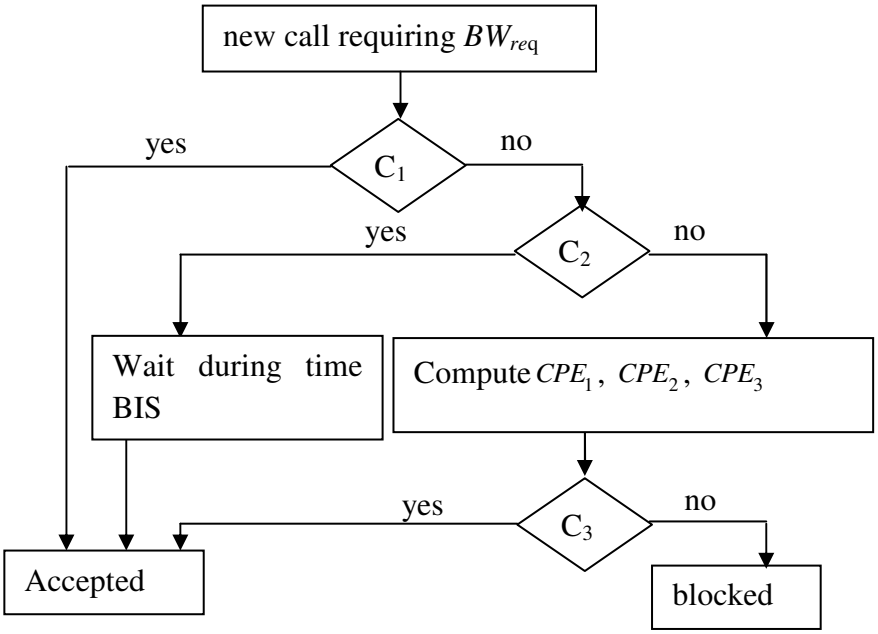
(1) The average of released bandwidth per call for calls that complete earlier than expected (CPE_1); the released bandwidth corresponds to the bandwidth passively reserved between the actual completion and the estimated completion.

(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., $CPE_1 + CPE_2 + CPE_3$) 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. Figure 16 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 ≤ BISt.
C_3: $CPE_1+CPE_2+CPE_3 \ge BWreq$

Figure 16: The operation of ECaC to accept or block a new call request.

To compute CPE_1, we use the database ECCD which 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) have the same type of day (e.g., weekend or weekdays) as the new call; the expression of CPE_1 is defined as follows:

$$CPE_1 = \frac{\sum_{l=1}^{n_w} bw_{alloc}^{l}}{n_w} \tag{4.24}$$

where bw_{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. CPE1 can 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(*t2* - *t1*) from ECCD **where** *cell ID* = *critical cell ID* **and** type_day(d) = type_day(current day)] $\geq (t_j^d - t_j^a)$".

To compute CPE_2 and CPE_3, we use the database UCD which 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 while the list of users required to compute CPE_3 is obtained 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 (resp. maximum) travel time, defined as [(distance to the *critical cell* border / maximum (resp. minimum) average velocity) + current time]), to reach the *critical cell* (resp. 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, $\left(t_j^a, t_j^d, c_j\right)$ be the user's estimated time window when the user will transit cell c_j, $\left(t_j^l, t_j^u, c_j\right)$ (where $t_j^u < t_j^a$) be the estimated time window when user g will enter cell c_j and $\left(t_f^l, t_f^u, c_f\right)$ (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 (resp. h) is said to "arrive late" (resp. "exit earlier") when his minimum (resp. maximum) travel time $t_{j,g}$ (resp. $t_{f,h}$) to reach the border of *critical cell* c_j is bigger (resp. smaller) than t_j^a; in other words, user g (resp. h) cannot reach (resp. exit), with

maximum (resp. minimum) average velocity, the *critical cell* c_j before (resp. after) the estimated arrival time of user u (see Figure 17).

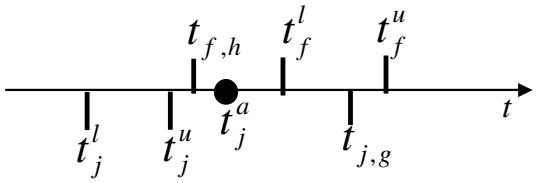


Figure 17: Illustration of "arrive late" and "exit earlier" users.

In the rest of this section, we refer to the reserved bandwidth for an incoming call or the allocated bandwidth to an outgoing call as required bandwidth.

Let L_g (resp. L_h) be the list of "arrive late" (resp. "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, $\left[t_j^a, t_j^d\right]$, 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:

$$CPE_2 = \sum_{g=1}^{n_g} bw_{req}^g \text{ and } CPE_3 = \sum_{h=1}^{n_h} bw_{req}^h \tag{4.25}$$

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 (1) the user is equal to x and (2) $t+d$ is bigger than t_j^a; the expression of bw_{req}^x is defined as follows:

$$bw_{req}^x = \sum_{l=1}^{n} bw(l) \tag{4.26}$$

where $bw(l)$ 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 SUM**(b) **from** UCD **where** $user = x$ **and** $t+d > t_j^a$".

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 handoff time estimation 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.

Figure 18 shows the architecture of MPBR including all databases and components together with their interactions.

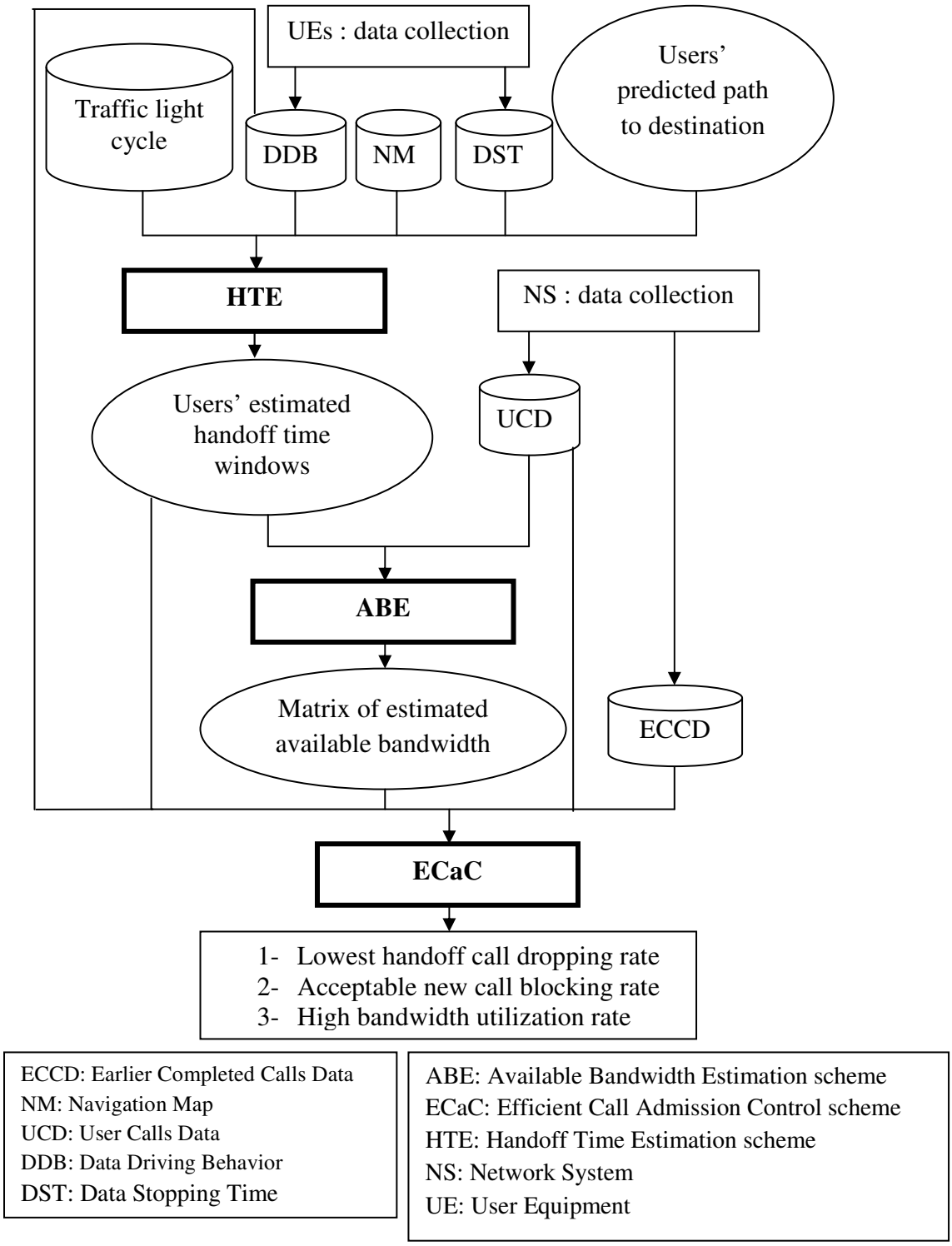


Figure 18: Illustration of MPBR processes.

4.4. Performance evaluation

In this section, we evaluate, via simulations, the performance of MPBR in terms of (a) handoff time prediction error; (b) new call blocking rate; (c) handoff call dropping rate; and (d) bandwidth utilization rate. We compare MPBR against the schemes described in [98], [97] and [28], 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 outperform existing approaches (e.g., [26, 27, 95, 101, 102]). 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.

4.4.1. Simulation setup

To evaluate MPBR, we used mobile user traces acquired from the Generic Mobility Simulation Framework (GMSF) project [130]. 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 in order 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 two-dimensional 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 Figure 19.

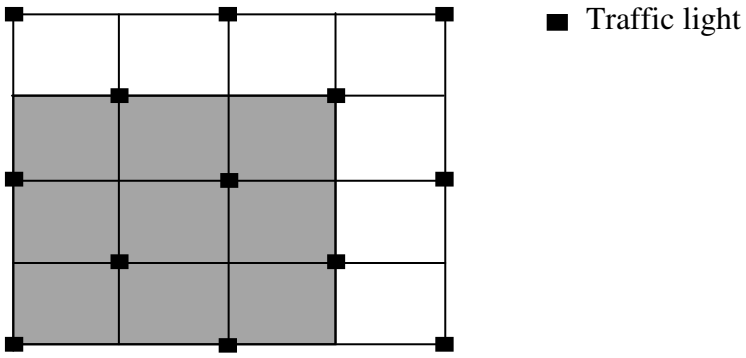


Figure 19: Cell coverage and traffic light locations.

The cellular structure can be typically seen in a metropolitan downtown area. We make the following assumptions for this two-dimensional environment: (1) each user has a predefined (predicted) path; (2) at the beginning of the simulations, each user u randomly chooses acceleration A_u (m/s^2) from within [0.1,0.2], deceleration D_u (m/s^2) from within [-0.2,-0.1], stopping time S_u (sec) from within [0,5], and maximum velocity Vm_u (m/s) from within [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, turn left or turn right according to his predefined path; (5) on each road segment, a user u reaches a maximum velocity Vm chosen randomly from within [Vm_u -1, Vm_u +1]; 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 seconds) to orange (5 seconds) and then to green (60 seconds).

Similar to [28, 86, 89, 101], new call requests are generated according to a Poisson distribution with rate λ (calls/second/user). The call time is assumed to be exponentially distributed with a mean of 300 sec. Table 12 shows the values of the parameters used in the simulations.

Table 12: 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} [97]	10 m/s

4.4.2. Results analysis

Simulation results are averaged over multiple runs with different pseudo random 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:

$$Average\_error = \frac{\sum_{u=1}^{q} \varepsilon_u}{q} \tag{4.27}$$

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*.

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

$$Rb = \frac{n_b}{m_b} \tag{4.28}$$

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).

- Handoff call dropping rate denoted by *Rd*; it is computed as follows:

$$Rd = \frac{n_d}{m_d} \tag{4.29}$$

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

- Bandwidth utilization rate denoted by *Rbw*; it is computed as follows:

$$Rbw = \frac{bw_{alloc}}{bw} \tag{4.30}$$

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.

Figure 20 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 0.03call/second/user, the cell capacity is set to 100 Mbps and the number of users in the simulation area is 1,500. We observe that, when BISt increases from 0 to 90sec, the average new call blocking rate decreases by 39.3% (i.e., average *Rb* at 0sec - average *Rb* at 90sec).

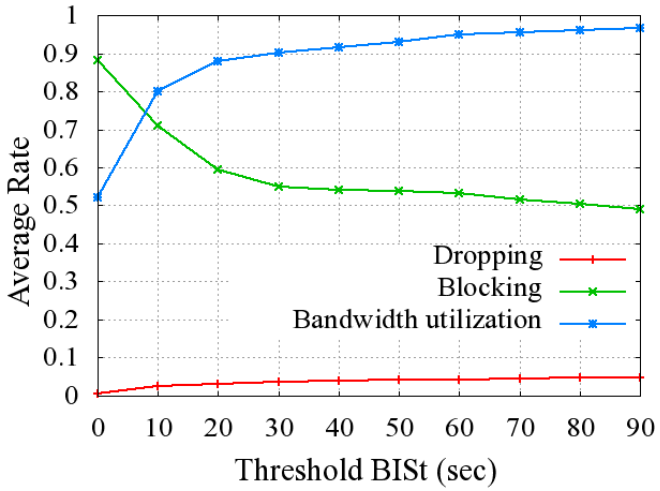


Figure 20: MPBR performance metrics (Rb, Rd, and Rbw) versus BISt variation.

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. 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 and thus the bandwidth utilization rate increases. We conveniently conclude that the BIS concept improves the performance of MPBR.

Figure 21 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.

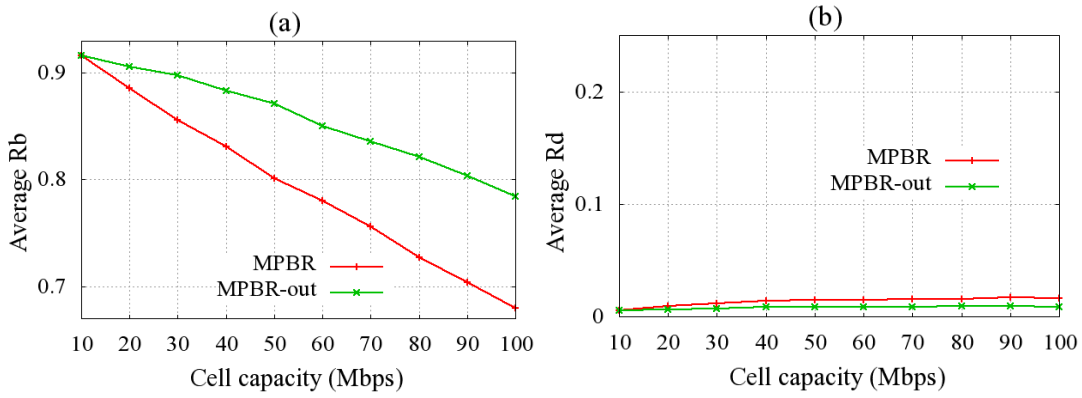


Figure 21: Impact of CPE on MPBR Performance metrics (Rb and Rd) versus cell capacity variation.

In this set of simulations, the call arrival rate λ is set to 0.03call/second/user, BISt is randomly chosen from {0, 30, 60, 90} and the number of users in the simulation area is 1,500. Figure 21a shows that MPBR outperforms MPBR-out; indeed, MPBR provides an average of 0.79 per 10 Mbps while MPBR-out provides an average of 0.86 per 10 Mbps; the average relative improvement (defined as [average *Rb* of variant - average *Rb* of MPBR]) of MPBR compared to MPBR-out is about 7% per 10 Mbps. Figure 21b shows that MPBR is slightly less efficient than MPBR-out: it provides an average of 0.008 per 10 Mbps while MPBR provides an average of 0.01 per 10 Mbps; the average relative improvement (defined as [average *Rb* of MPBR - average *Rb* of variant]) of MPBR-out compared to MPBR is about 0.2% 10 Mbps which is negligible. Thus, we conclude that MPBR provides a reduction of 7% 10 Mbps of new call blocking rate with negligible increase in handoff call dropping rate.

Figure 22 shows the average prediction error gap (computed by Equation 4.27) 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 to 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 under-saturated condition (from 501 to 1,000 users), and about 35.3% per 150 users in the congested condition (from 1,001 to 1,500 users). Overall, the average relative improvement of MPBR compared to 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 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). Figure 22 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 and, thus, prediction based on the speeds of all previous users provides better performance (compared to 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 to the case of a small number of users).

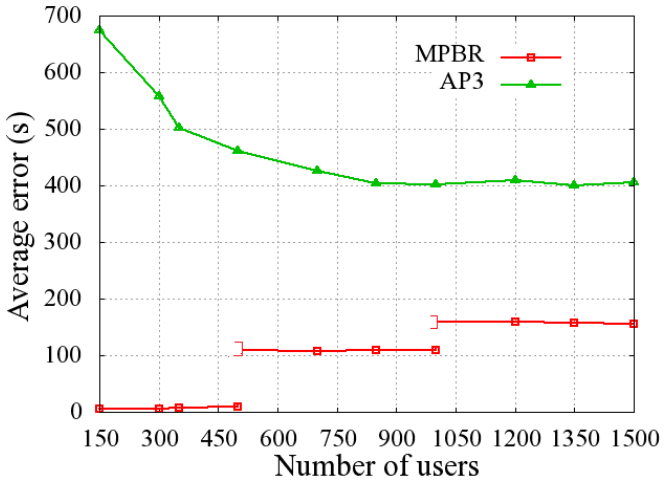


Figure 22: Average prediction error gap versus number of users.

Figure 23 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. Figure 23a 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 0.73 per 10 Mbps while MPBR provides an average call blocking rate of 0.79 per 10 Mbps; thus, the average relative improvement of AP3 compared to MPBR is about 6% per 10 Mbps. 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. Figure 23b shows that MPBR outperforms AP1, AP2 and AP3. For example, MPBR provides an average of 0.01 per 10 Mbps while AP1 (slightly more efficient than AP2 and AP3 in this scenario) provides an average handoff call dropping rate of 0.52 per 10 Mbps; overall, the average relative improvement of MPBR compared to AP1 is about 51% per 10 Mbps. 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. Even though 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).

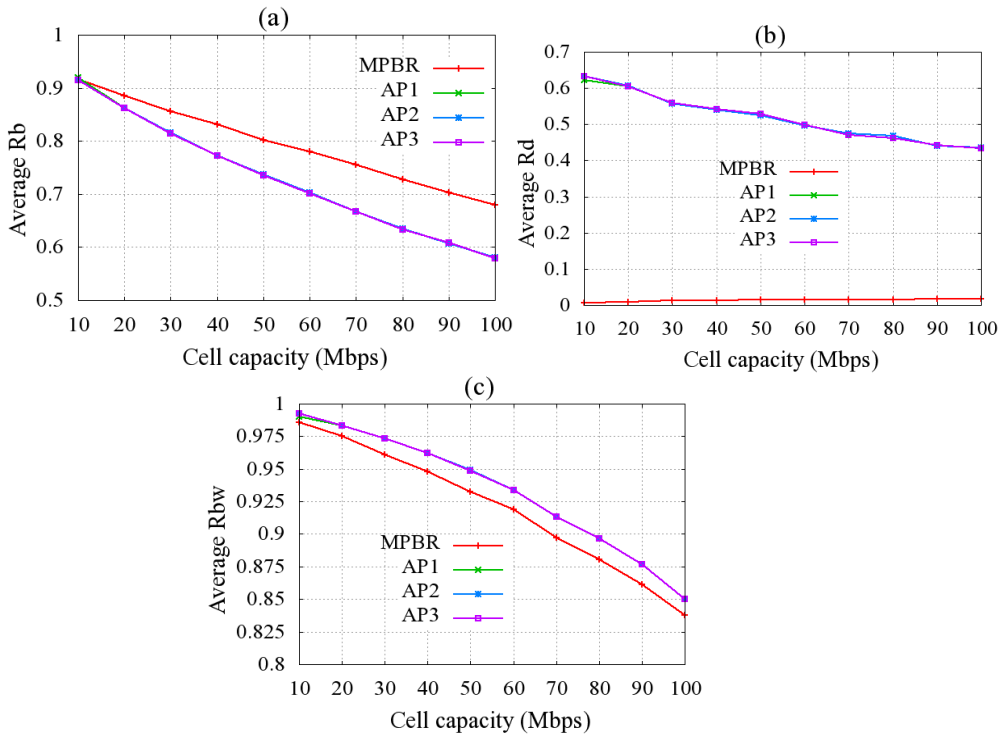


Figure 23: Performance metrics (Rb, Rd, and Rbw) versus cell capacity variation.

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 Figure 22) along the user path to destination before the acceptance of the call. Even though AP1 uses similar reservation mechanism (i.e., reservation along user's path to destination), it does not make use of efficient available bandwidth estimation scheme; nonetheless, AP1 slightly outperforms AP2 and AP3 in this scenario (Figure 23b). Figure 23c 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. Figure 23c 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 0.93 per 10 Mbps while MPBR provides an average bandwidth

utilization rate of 0.92 per 10 Mbps; the average relative improvement of AP3 compared to MPBR is about 1% per 10 Mbps which is negligible. We conclude that, compared to AP1, AP2 and AP3, MPBR provides a considerable reduction of 51% per 10 Mbps in handoff call dropping rate and a slight increase of 6% per 10 Mbps in new call blocking rate with similar bandwidth utilization irrespective of the network cell capacities. The 6% new call blocking rate increase is a small price to pay for the small handoff call dropping rate.

Figure 24 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 100Mbps, BISt is chosen from within the set {0, 30, 60, 90} sec with equal probability and the number of users in the simulation area remains 1,500. Figure 24a 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 0.62 per 0.01 call arrival rate while MPBR provides an average new call blocking rate of 0.73 per 0.01 call arrival rate; the average relative improvement of AP3 compared to MPBR is about 11% per 0.01 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. Figure 24b shows that MPBR outperforms AP1, AP2 and AP3; MPBR provides an average handoff call dropping rate of 0.02 per 0.01 call arrival rate while AP2 (slightly more efficient than AP1 and AP3 in this scenario) provides an average handoff call dropping rate of 0.54 per 0.01 call arrival rate. Overall, the average relative improvement of MPBR compared to AP2 is about 52% per 0.01 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. Even though 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. 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 Figure 22) along the user path to destination before the acceptance of the call. Figure 24c 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 0.89 per 0.01 call arrival rate while MPBR provides an average of 0.85 per 0.01 call arrival rate; the average relative improvement of AP3 compared to MPBR is about 4% per 0.01 call arrival rate; which is still negligible. We conclude that, compared to AP1, AP2 and AP3, MPBR provides a considerable reduction of 52% per 0.01 call arrival rate in handoff call dropping rate and an increase of 11% per 0.01 call arrival rate in new call blocking rate with similar bandwidth utilization irrespective of call arrival rates. The 11% new call blocking rate increase is a small price to pay for the small handoff call dropping rate.

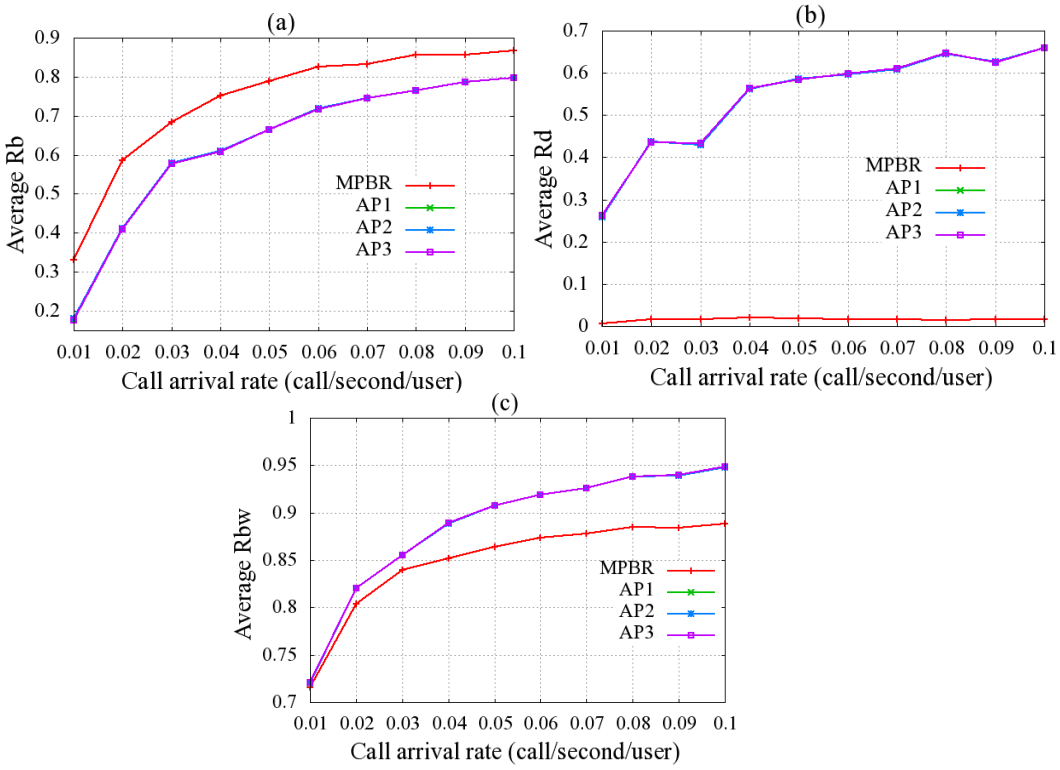


Figure 24: Performance metrics (Rb, Rd and Rbw) versus call arrival rate variation.

4.5. Conclusion

In this paper, we 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) a handoff time estimation 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) efficient call admission control scheme that aims to control bandwidth allocation in cells in order 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 [97, 98] and one closely related scheme [28]. 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 ANDSF (Access Network Node Discovery Function) node [131-133], using a system design similar to that proposed by the authors in [134], and envisioning similar additional components to ANDSF. Indeed, ANDSF can be used to (a) collect mobility features from UEs, using in turn HTE; (b) predict available bandwidth of mobile backhaul along the predicted trajectory of user using time series as in [135] and ABE proposed in this paper; and (c) 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).

Chapitre 5 :

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

Apollinaire Nadembéga, Abdelhakim Hafid, Tarik Taleb

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 (QoS) and ensure steady Quality of Experience (QoE) for users of mobile multimedia streaming services. Indeed, the support of uniform streaming rate during the entire course of a streaming service while the user is on the move is a challenging issue. In this paper, we propose a framework, together with schemes, that 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, 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.

Status: This article is submitted to IEEE Transactions on Wireless Communications 2013; it is based on the following published papers:

A. Nadembéga, A. Hafid and T. Taleb. *A Framework for Mobility Prediction and High Bandwidth Utilization to Support Mobile Multimedia Streaming,* IEEE ANTS, Chennai, India, Dec. 2013.

T. Taleb, A. Hafid, and A. Nadembéga. *Mobility-Aware Streaming Rate Recommendation System*, IEEE Globecom, Houston, Texas, USA, Nov. 2011.

5.1. 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. 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 [16]). 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 [21, 30, 31]. The main limitation of these schemes [21, 30, 31] 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 [21, 30, 31, 136], making them scale with the number of users, by developing (1) an Aggregate Path Prediction Model called APPM; (2) an Aggregate Handoff Times Estimation Scheme called AHTES; and (3) 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 [30].

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 behaviour 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.

The remainder of this paper is structured as follows. Section 5.2 presents some related work. Section 5.3 describes the proposed framework along with the envisioned mobile network architecture. Section 5.4 evaluates the performance of the proposed framework and showcases its potential in achieving its design objectives. Section 5.5 concludes this paper.

5.2. Related Work

CAC and bandwidth management schemes can be classified into two categories: (a) non-predictive schemes [14, 27, 94, 97, 98, 102-106, 137] (based only on the source cell information) and (b) predictive schemes [28, 86, 89, 91, 93, 95, 96, 100, 101, 114] (based on the mobility information of a mobile user). 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 [26-28, 86, 89, 91-97] and two or more cells [98, 114]) and (b) the way handoff

requests are handled (e.g., non-prioritized or prioritized handoff [26-28, 86, 89, 91-93, 95, 97-103, 114, 137]). According to [16], prioritized handoff schemes which are distributed and predictive are those which better satisfy bandwidth requirements of users from source to destination.

In the following, we briefly overview representative schemes [86, 91, 114] that are closely related to our proposed approach (i.e., predictive mobile-oriented schemes). In these three schemes [86, 91, 114], a handoff call is admitted if there is enough available bandwidth in the next cell; otherwise, it is dropped. Dias *et al.*[86] 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 periodic exchange of load estimates that are likely to move or migrate from one cell to its neighbors, Dias *et al.* predict available bandwidth of next cell. This scheme [86] suffers from three key limitations: (1) aggregation is used only to estimate available bandwidth which negatively impacts the scalability of the scheme; (2) available bandwidth estimation is not predictive (i.e., based on the behavior/state of network cells [16, 138]); 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 [28, 91, 94, 101, 114]; and (3) 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.*[91] 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 $[BCc_{i+1} - (BEc_{i+1} + BRc_{i+1})]$ where BCc_{i+1} (resp. BEc_{i+1}/BRc_{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.* [114] 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 [114] 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: (1) They rely on the current behavior/state of the network cells [17, 98] 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; (2) The schemes that make use of mobility prediction techniques either do not take into account users' aggregation [27, 86, 101, 107-113], require additional equipment [26, 27], generate significant traffic overhead in terms of mobility data exchanges between users and network backbone [49], do not consider stopping times [17, 18, 20, 26-28, 49], make use of old road traffic data [18, 28] or rely only on historical data about previous users [28]; (3) admission control procedures are limited to source cell and possibly also next cell [26-28, 86, 91, 97, 114]; and/or (4) they rely only on historical network bandwidth observations or users transit times in cells [86, 91, 98, 137]. In this paper, we propose a scheme, to process call requests, that proposes solutions to these limitations.

5.3. 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 (1) the mobile user path to destination; (2) the entry/exit times of the mobile user to/from cells along the path to destination; and (3) 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. Table 13 shows the list of symbols/variables that are used to describe the proposed schemes.

Table 13: Summary of notations.

Symbol	Description	Symbol	Description
D	Boundary density value of subzone	$PBP(P_i,x,y)$	Profile-based path from *DSZ x* to *DSZ y* of road profile P_i
$R_{Hi,}$	Preference rate of road segments characterized by *Hi*	$p(s,P_i)$	Probability that adjacent road segment s is the next road segment towards destination DSZ, given road profile P_i
R_{Ij}	Preference rate of road intersection characterized by *Ij*	$D(Z,t)$	Density of subzone Z at time t
r_H	Preference threshold of road segment	L_l	Length of portion l of sub-road segment already transited by a user
r_I	Preference threshold of road intersection	$n(t)$	Number of users on road segment at time instant t
V_{Hi}	Preference of road segment characterized by *Hi*	$d(t)$	Density of road segment at time instant t
V_{Ij}	Preference of road intersection characterized by *Ij*	d_1	Lower boundary density values of road segment
$r(\theta_s)$	Deviation rate of each adjacent road segment	d_2	Upper boundary density values of road segment
Ψ	Pre-selected adjacent road segments	$u\Delta T^R$	Transit time on sub-road segment R by user u who has already transited R
N_a	Cardinality of Ψ	ΔT_R^u	Transit time that will be required to transit R by user u who is currently on sub-road segment R
L_R	Length of sub-road segment R	$u\Delta S^j$	Transit time of road connection zone J by user u who has already transited J
L	Length of road segment	$F_{SiGkSa}()$	CDF of transit/travel times of link *SiGkSa*
E_B^A	Selection area from A towards B	ΔT_{SiGkSa}	Transit time of link *SiGkSa*

In the following section, after stating our assumptions, we present the aggregate path prediction model (APPM) and the aggregate handoff times estimation scheme (AHTES). Then, we present the architecture of IPMBRF. Finally, we present the IPMBRF operations in processing new call requests.

5.3.1. 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*→*b*) is different from the navigation direction *b* towards *a* (*b*→*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:

$$D(Z,t) = \frac{Num_u(Z,t)}{A_Z} \tag{5.1}$$

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).

5.3.2. Definitions

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

5.3.2.1. 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).

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.

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.

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

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.

Navigation Map Register (NMR): NMR is a database; NMR 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.

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.

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.

5.3.2.2. 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. We assume that each road segment S_i (e.g., $a \rightarrow b$ shown in Figure 25) is divided into sub-road segments where *SiGk* is the k^{th} sub-road segment (see Figure 25);

Road segment portion: a road segment portion is the road between two sub-road segments of the same road segment; for example, in Figure 25, we refer to the road between sub-road segments *SiGk* and *SiGw* as the road segment portion *SiGkGw*.

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 Figure 25, 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 Figure 25).

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 Figure 25, 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*).

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 Figure 25 the path portion that represents road between *SiGk* and *S1G4* consists of link *SiGkS1* and road segment portion *S1G1G4*.

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).

Data of Driving Behaviors Register (DDBR): DDBR is a database; whenever a user *UID* transits sub-road segment *SID* with an acceleration *a* and a 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).

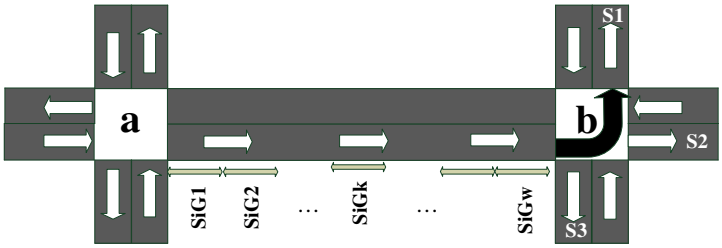


Figure 25: Illustration of sub-road segments

5.3.3. 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; 2^n 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 [21].

5.3.3.1. 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 2^n 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 L_1. Making use of list L_1, we compute the pair (Hi, d_{Hi}) and store it in list L_2; d_{Hi} represents the total length of road segments which are already transited by the user and have characteristic H_i. L_2 can be easily computed using the following SQL query:

Select *Hi*, **SUM**(*l*) **as** d_{Hi} **from** L_1 **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:

$$R_{Hi} = \frac{d_{Hi}}{\sum_{l=1}^{k} d_{Hl}} \tag{5.2}$$

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 L_3. Making use of list L_3, we compute the pair (I_j, n_{Ij}) and store it in list L_4; n_{Ij} represents the total number of times the user has transited road intersections characterized by I_j. L_4 can be easily computed using the following SQL query:

Select *Ij*, **COUNT**(*) **as** n_{Ij} **from** L_3 **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:

$$R_{Ij} = \frac{n_{Ij}}{\sum_{i=1}^{m} n_{Ii}} \tag{5.3}$$

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:

$$r_H = \frac{1}{k} (\text{resp. } r_I = \frac{1}{m}) \tag{5.4}$$

When $R_{Hi} \geq r_H$ (resp. $R_{Ij} \geq r_I$), 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:

$$V_{Hi} = R_{Hi} \geq r_H (\text{resp.} V_{Ij} = R_{Ij} \geq r_I) \tag{5.5}$$

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

$$P = \bigwedge_{i=1}^{k} V_{Hi} \bigwedge_{j=1}^{m} V_{Ij} \tag{5.6}$$

5.3.3.2. Generation of profile-based path and users' path prediction

For each road profile *Pi*, APPM generates a profile based path, *PBP(P_i,x,y),* from source DSZ *x* to destination DSZ *y* that best meets road profile P_i. 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 $E_{destination\_DSZ}^{current\_location}$. 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 *s* are those that belong to the following set:

$$\Psi = \bigcup_s \left\{ s \,\middle|\, r(\theta_s) \geq r(45^o) = 3/4 \right\} \tag{5.7}$$

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:

$$p(s, P_i) = \frac{Num(s, P_i)}{\sum_{a=1}^{N_a} Num(a, P_i)} \tag{5.8}$$

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 the list 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 $O(n*e*g^2)$.

5.3.4. 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.

5.3.4.1. Traffic flow and queuing models

5.3.4.1.1. 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:

$$f(t) = d(t) \times v(t) \tag{5.9}$$

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:

$$d(t) = \frac{n(t)}{L} \tag{5.10}$$

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:

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.

Model for differences in driving behavior: in this paper, driving behavior is based on the velocity model proposed in [58]; indeed, driving behavior is a cycle of acceleration, drive at a constant velocity, deceleration and finally stopping.

Stationarity of traffic: During each estimation interval, the parameters of traffic light cycles are constant; i.e., the time 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".

5.3.4.1.2. 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* ($d(t) \le d_1$); (ii) *under-saturated conditions* ($d_1 < d(t) < d_2$) and (iii) *congested conditions* ($d(t) \ge d_2$).

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:

$$u\Delta T^R = uT_e^R - uT_a^R \tag{5.11}$$

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:

$$\Delta T_R^u = L_R \times \frac{u\Delta T^l}{L_l} \tag{5.12}$$

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 Equation (5.11). We also define the expression of transit time of road connection zone J by user u (who has already transited J) as follows:

$$u\Delta S^J = uS_e^J - uS_a^J \tag{5.13}$$

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.

5.3.4.2. 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 [31].

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 Equation (5.10). The final output of AHTES, for a given user request, is an *n*-tuple: $\Omega = \langle (t_1, c_1), ..., (t_i, c_i), ..., (t_n, c_n) \rangle$ where t_i is the value of the estimated time when the current user will reach cell C_i and $C_1, \ldots, C_n$ 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 (1) users who have already transited *SiGk* during the last 30 minutes; these times are computed based on Equation (5.11), and (2) users who are currently on *SiGk*; these times are computed based on Equation (5.12). 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:

$$v_{SiGk}(\Delta T_u) = \frac{n_{SiGk}^{\Delta T_u}}{n_{SiGk}} \tag{5.14}$$

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 Equation (5.13). 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:

$$w_{SiSa}(\Delta S_u) = \frac{n_{SiSa}^{\Delta S_u}}{n_{SiSa}} \tag{5.15}$$

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 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:

$$\rho_{SiGkSa}(\Delta T) = \left(\left[(((v_{SiGk} * v_{SiGk+1}) * v_{SiGk+2}) * ...) * v_{SiGw} \right] * w_{SiSa} \right)(\Delta T) \tag{5.16}$$

Using Equation (5.16), we derive the cumulative distribution function, F_{SiGkSa}, of transit times of link *SiGkSa* as follows:

$$F_{SiGkSa}(\Delta T') = \sum \rho_{SiGkSa}(\Delta T \leq \Delta T') \tag{5.17}$$

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:

$$\Delta t_{SiGkSa} = F^{-1}_{SiGkSa}(0.5) \tag{5.18}$$

Thus, based on CDF of transit times of each link *SiGkSa* of current road segment *Si*, we obtain its transit time table of *Si*. Using the example shown in Figure 25, we compute the transit time table of *Si* (see Table 14). AHTES computes a table (Table 14) 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 *Si*, 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 *Si*; f is also the number of links of road segment *Si*; thus, adding $O(f)$ complexity for one road segment; for g road segments, AHTES adds $O(f*g)$ complexity.

Table 14: Transit time table of road segment *Si.*

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}

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:

$$t_i = t_0 + \Delta t_{S1G3S3} + \Delta t_{S3G1S4} + (\Delta t_{S4G1S5} - \Delta t_{S4G4S5}) \quad (5.19)$$

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.

5.3.5. Architecture

Figure 26 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 Figure 26) allows cell towers and gateways to be inter-connected. Figure 27 shows the architecture of IPMBRF whose operation is performed by User Equipment (UE) and the Controller (CTL) which is located in the network system (NS).

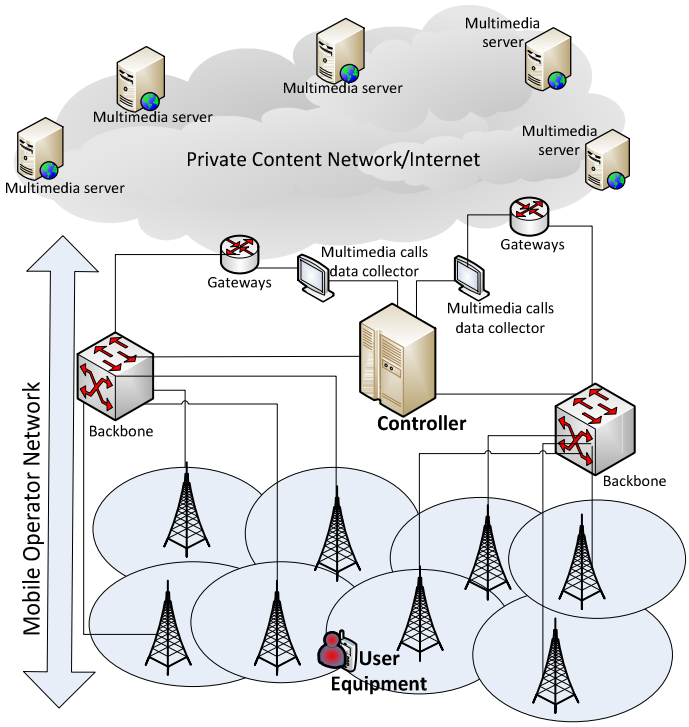


Figure 26: Envisioned network architecture.

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). Figure 27 shows CTL that consists of four main modules namely Path Predictor-Network side (PP-N), Handoff 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.

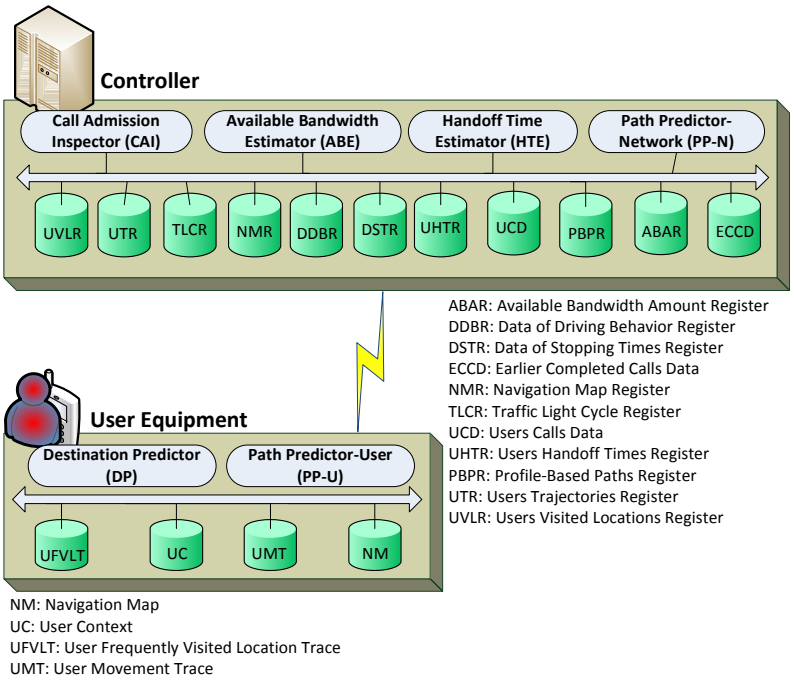


Figure 27: IPMBRF Architecture.

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 [21, 30, 31, 136].

5.3.6. 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. Figure 28 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 [30]; 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 Equation (5.1); 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 [21]), the predicted path and sends it back to CTL; otherwise, determines, by using PP-N (implements APPM; see Section 5.3.3). 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 [31] (resp. AHTES, see Section 5.3.4) 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.

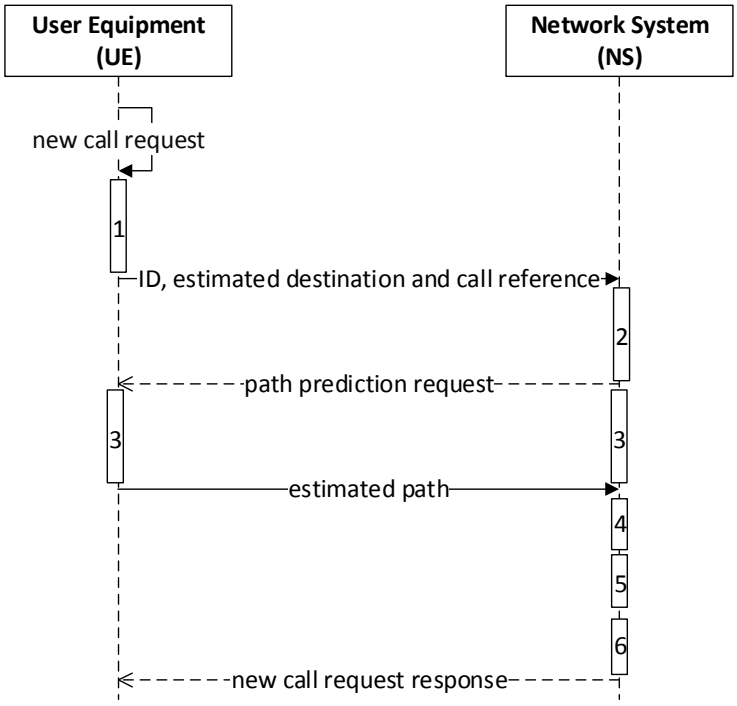


Figure 28: IPMBRF process for new call acceptance/rejection.

5.4. Performance evaluation

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

5.4.1. APPM performance evaluation

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

$$A_p(E_{act}, E_{pred}) = \frac{2.\left|E_{act} \cap E_{pred}\right|}{\left|E_{act}\right| + \left|E_{pred}\right|} \tag{5.20}$$

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. C_t is defined as follows:

$$C_t = \sum_{i=1}^{N_U} T_i \tag{5.21}$$

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

We compare the performance of APPM against PPM described in [21]. Indeed, APPM proposes an aggregate path prediction model while PPM proposes an individual path prediction model. Simulation results are averaged over multiple 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 Equation (5.20) (resp. Equation 5.21); thus, to obtain the simulation results shown in Figure 29 we compute the average of the *five hundred* runs.

5.4.1.1. Simulation setup

To evaluate APPM and PPM, we used real mobile user traces, acquired from the MIT media laboratory's database available in the context of the Reality Mining Project [74]. A subject trace consists of a sequence of locations; a location contains user ID, date, cell ID, arrival time to the cell, and departure time from the cell. The mobile users, in this project, consist of students and staff at a major university during ten months. We make the following assumptions: (1) the dense zones (DSZs) refer to the blocks of cells where more than 20% of subjects are located during specific times of the day; we define four specific times: home time (e.g., within [10h p.m. and 8h a.m.]), working time (e.g., within [8h a.m. and 1h p.m.] and [2h p.m. and 6h p.m.]), restaurant time (e.g., within [1h p.m. and 2h p.m.]); leisure time (e.g., within [6h p.m. and 10h p.m.]); (2) the subjects who have 50% of path similarity during the length of the learning phase have the same road profile; we identify seventeen road profiles; (3) the current location (where the prediction process is executed) corresponds to 10% of the path from trip origin to destination; (4) the length of the learning phase is 70% of the considered data collection (i.e., 70 days) and (5) prediction length d_t is 30 minutes. The values of the simulation parameters used by DPM [30] are: f_{th}=1/30; θt_{h1}=90^o and θt_{h2}=45^o.

5.4.1.2. Results analysis

Figure 29a shows that PPM slightly outperforms APPM; indeed, PPM provides an average accuracy of 0.90 per user while APPM provides an average accuracy of 0.86 per user; thus, the average relative improvement (defined as [average *Ap* of PPM - average *Ap* of APPM]) of PPM compared to APPM is about 4%. 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.

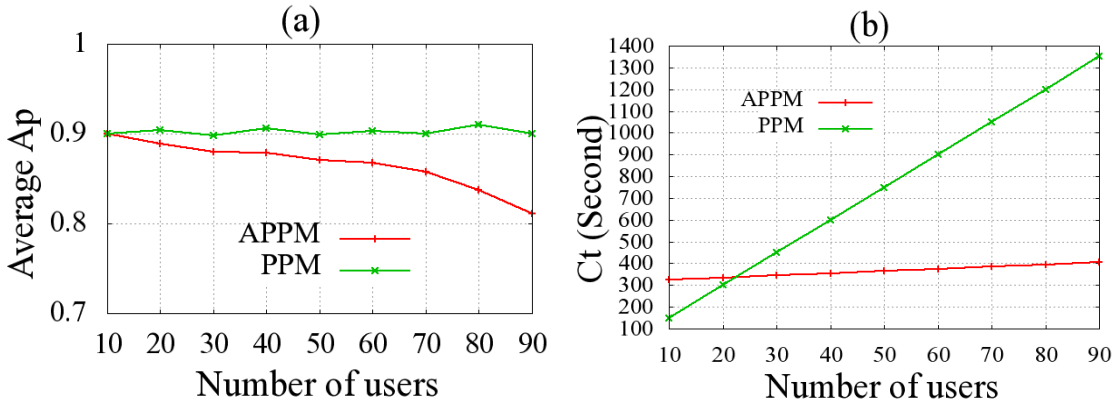


Figure 29: The performance of APPM and PPM

Figure 29b shows that APPM outperforms PPM. APPM provides an average time complexity of 7.28 per user while PPM provides an average time complexity of 15 per user; thus, the average relative improvement (defined as [average *Ct* of PPM - average *Ct* of APPM]) of APPM compared to PPM is about 51.47%. We observe that APPM time complexity increases slowly while PPM time complexity increases rapidly. Indeed, the computation time of PBPs tables of APPM, for the twenty-nine DSZs and the seventeen road profiles, is about 314 seconds; the predicted path computation of APPM (without considering PBPs tables computation) for one user is about 1 second while the predicted path computation of PPM for one user is 15 second.

We conclude that, compared to PPM, APPM provides a reduction of 51.47% in time complexity and a slight decrease of 4% in accuracy; the 4% accuracy decrease is a very small price to pay for the small time complexity.

5.4.2. 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 *Rb,* is computed as follows:

$$Rb = \frac{n_b}{m_b} \tag{5.22}$$

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 *Rd*, is computed as follows:

$$Rd = \frac{n_d}{m_d} \tag{5.23}$$

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 *Ebw*, is computed as follows:

$$Ebw = \frac{|BWa - BWe|}{BWa} \tag{5.24}$$

where *BWa* is the actual available bandwidth and *BWe* is the estimated available bandwidth; *BWa* and *BWe* are measured whenever a call is blocked or dropped.

We compare the performance of IPMBRF against the schemes described in [114] and [91], 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

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 *Rb* (resp. *Rd/Ebw*) using Equation (5.22) (resp. Equation 5.23/5.24); thus, to obtain the simulation results shown in Figure 30 we compute the average of the *one thousand* runs.

5.4.2.1. Simulation setup

To evaluate IPMBRF, we used mobile user traces acquired from the Generic Mobility Simulation Framework (GMSF) project [130]; 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 [89], 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 the road segment has a traffic light. The cellular structure can typically be seen in a metropolitan downtown area. We make the following assumptions for this two-dimensional environment: (1) each user has a predicted path with 86% of accuracy; this path prediction accuracy is the average accuracy which is obtained after the simulation of APPM (see Figure 29a); (2) at the beginning of the simulations, each user u randomly chooses acceleration A_u (m/s^2) from within [0.1,0.2], deceleration D_u (m/s^2) from within [-0.2,-0.1], stopping time S_u (sec) from within [0,5], and maximum velocity Vm_u (m/s) from within [10,14]; (3) after each stop, the initial velocity is 0; (4) at the intersection of two road segments, a user randomly selects to continue straight, turn left or turn right; (5) on each road segment, a user u reaches a maximum velocity Vm chosen randomly from within [Vm_u -1,Vm_u +1]; 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 about 900m [28]; (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 seconds) to orange (5 seconds) and then to green (60 seconds) [58].

Similar to [28, 86, 89, 101], 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) [28, 86, 102, 106]; in the simulations, we focus on the improvement of handoff call dropping rate; thus, similar to [28, 86, 89, 97, 98, 101], 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. 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=400m, d_1=0.025 user/m; d_2=0.075user/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 $\lambda$is 0.03call/second/user; and the cell capacity is 100BUs.

5.4.2.2. Results analysis

Figure 30 shows (a) the average new call blocking rate and (b) the average handoff call dropping rate for different available bandwidth estimation errors. Figure 30a 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 *Ebw* while IPMBRF provides an average call blocking rate of 0.42 per 5% of *Ebw*; thus, the average relative improvement (defined as [average *Rb* of IPMBRF - average *Rb* of AP2]) of AP2 compared to IPMBRF is about 17% per 5% of *Ebw*. We observe that, for the three schemes, the average new call blocking rate decreases when *Ebw* increases. This is expected since when *Ebw* 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. Figure 30b shows that IPMBRF outperforms AP1 and AP2; IPMBRF provides an average handoff call dropping rate of 0.16 per 5% of *Ebw* while AP1 (slightly more efficient than AP2 in this scenario) provides an average handoff call dropping rate of 0.65 per 5% of *Ebw*; overall, the average relative improvement (defined as [average *Rd* of AP1 - average *Rd* of IPMBRF]) of IPMBRF compared to AP1 is about 49% per 5% of *Ebw*.

We observe that, for the three schemes, the average handoff call dropping rate increases with *Ebw*. This is expected since when *Ebw* 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 *Ebw*, 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 (Figure 30b) 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.

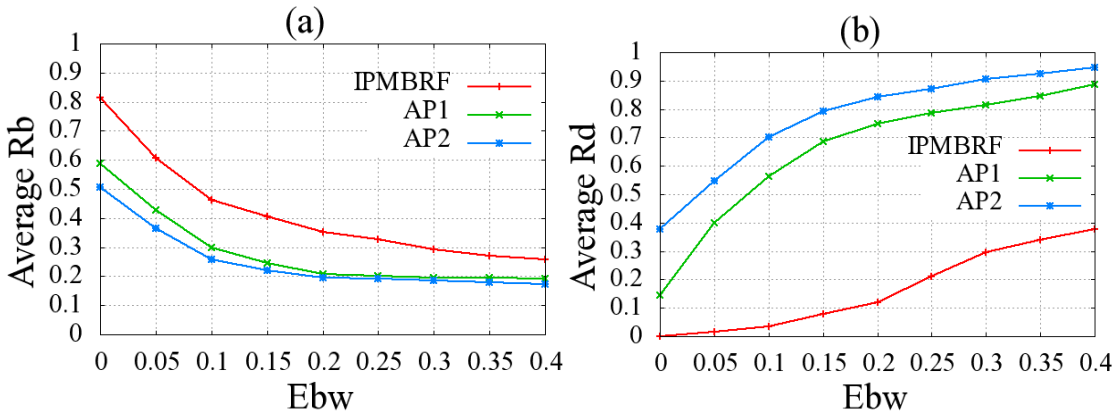


Figure 30: Rb and Rd versus Ebw

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

5.5. 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 [91, 114]. 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: (1) 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; (2) 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 (3) 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.

Chapitre 6 : Conclusion et travaux futurs

Dans ce chapitre, nous passerons en revue les contributions et les résultats de cette thèse. Nous dresserons aussi les perspectives pour des travaux de recherche futurs.

6.1. Contributions et résultats de la thèse

La prédiction de la mobilité des utilisateurs, le contrôle d'admission et la réservation de la bande passante dans les WCNS ont constitué les trois principaux sujets de cette thèse.

Dans la première contribution, présentée au chapitre 3, nous nous sommes intéressés à l'amélioration de la précision du chemin vers une destination prédite à l'avance. Pour l'obtention de la destination à l'avance, nous avons proposé un modèle de prédiction de la destination. L'amélioration de la prédiction du chemin est liée (1) au type de données utilisé ; et (2) à l'usage de certains critères. Les types de données pouvant être utilisés sont les données historiques de déplacements des utilisateurs ou la connaissance des goûts et des activités des utilisateurs. Quant aux critères, ceux sont : l'origine du déplacement, la position précédente, la position courante, la direction du déplacement, la segmentation, la position des utilisateurs dans le voisinage, la vitesse, la distance entre la position courante et la prochaine intersection de routes, la distance entre la position courante et la prochaine cellule, le contexte de l'utilisateur, le comportement collectif des utilisateurs, le jour de la semaine et le moment de la journée. Cependant, la question principale est : quels critères utilisés, comment les utiliser et comment déterminer la meilleure précision avec les critères choisis? Alors, nous avons exploré la capacité de l'usage simultané des deux types de données cités plus haut et la combinaison de plusieurs critères à produire une prédiction de la mobilité plus précise. Ces types de données ont déjà été utilisés par plusieurs travaux mentionnés dans la revue de littérature, cependant aucun d'entre eux n'a proposé un usage simultané des deux types. L'avantage d'une telle initiative est de bénéficier des atouts de chaque type de données et aussi de réduire les effets de leurs limites sur la précision. Notre contribution consistait à proposer une approche probabiliste de prédiction de la mobilité basée sur l'apprentissage du comportement de l'utilisateur en termes de déplacements. Plus précisément, nous avons proposé une extension du *second ordre de Markov* en utilisant les données historiques des

déplacements antérieurs des utilisateurs. Nous l'avons combiné à la théorie de l'évidence avec les données relatives à la connaissance des goûts et des activités des utilisateurs. Sachant que l'usage des données historiques demeure le meilleur moyen de prédiction, l'intuition derrière l'usage de la connaissance de l'utilisateur est de limiter les erreurs de prédiction liées à l'usage des données historiques. En nous servant de ces deux types de données, nous avons conçu des algorithmes de prédiction qui combinent les meilleurs critères cités plus haut. La combinaison des critères a été un élément clé de cette contribution en plus de l'usage des deux types de données. Spécifiquement, en nous servant des données historiques, des données courantes et de la connaissance de l'utilisateur, nous avons sélectionné un bloc de lieux comme étant la destination de l'utilisateur ; les données historiques font référence aux fréquences de visite de certains lieux tandis que les données courantes font référence aux directions courantes des utilisateurs ; une direction courante représente la direction de l'origine du déplacement vers la position courante ; la connaissance de l'utilisateur fait référence à ses goûts et ses activités. La formation des blocs s'est basée sur les lieux déjà fréquentés par l'utilisateur qui peuvent être atteints en utilisant un même chemin pendant une durée de déplacement prédéfinie. L'idée de la formation des blocs a permis de réduire le nombre de destinations possibles et d'accroitre la précision sur le choix de la destination. L'utilisation de la direction courante a permis de faire une présélection des lieux situés dans la direction du déplacement comme étant les potentielles destinations. C'est également une technique qui a contribué à en compte les informations courantes sur le déplacement de l'utilisateur. L'avantage derrière cette idée est de réduire l'espace de recherche de la destination. Ce qui réduit aussi la charge de traitement du processus de sélection de la destination. Puis, sachant le bloc de destinations et le chemin parcouru depuis l'origine du déplacement, nous avons proposé un modèle de prédiction du chemin de l'utilisateur vers le bloc de destinations. Plus précisément, à chaque intersection de routes, en faisant usage de la structure spatiale de la zone de déplacement, des données historiques (fréquence de routes utilisée), du bloc de destinations estimées et des données courantes (chemin déjà sélectionné ; c'est-à-dire la séquence de segments de route sélectionnés de l'origine du déplacement vers la position courante et la direction future ; c'est-à-dire la direction de la position courante vers le bloc de destinations estimées), nous avons évalué la possibilité d'être choisi de chaque segment de route adjacent au dernier segment de route sélectionné. La direction future a aidé à donner plus de chance aux segments de route orientés

dans la même direction que le bloc de destinations estimées. La structure spatiale de la zone de déplacement a contribué à éliminer les segments de route adjacents au dernier segment de route sélectionné qui ne permettent pas d'atteindre le bloc de destinations estimées. Ainsi, la direction future et la structure spatiale de la zone de déplacement nous ont permis d'améliorer la précision de la prédiction. Il faut noter également que nous avons filtré les données historiques selon le type de jours (jour de travail, jour de repos, fin de semaine) et le moment de la journée (matinée, midi, après-midi, soirée, nuit) avant leur usage. L'idée derrière l'usage de cette procédure de filtrage est de prendre en compte les habitudes des utilisateurs. En somme, la combinaison des deux types de données, le choix et la méthode de manipulation des critères que nous avons exploités pour la conception de notre modèle de prédiction de la mobilité ont amélioré de façon significative la précision de la prédiction du chemin des utilisateurs.

La deuxième et la troisième contribution ont été focalisées sur le support de la QdS. Plus spécifiquement, elles se sont intéressées au support du transfert intercellulaire sans l'arrêt forcé des sessions en cours dans les WCNs. En d'autres termes, elles ont été dédiées au support de la disponibilité de la bande passante pour les sessions des applications et services en cours d'exécution durant le déplacement des utilisateurs. Toutefois, nous nous sommes assurés de ne pas causer une augmentation du taux de refus des demandes de nouvelles sessions. En résumé, notre objectif dans ces deux contributions est de fournir un taux d'arrêt forcé et inopiné des sessions en cours quasi nul tout en maintenant un taux acceptable de refus des demandes de nouvelles sessions. L'objectif visé est de permettre aussi un taux d'utilisation élevé de la bande passante disponible. L'approche générale adoptée dans cette thèse a été celle de la réservation de la bande passante le long des chemins des utilisateurs (séquence de segments de route à visiter). Pour cette approche, nous devons savoir à priori la quantité de bande passante qui sera disponible dans les cellules au moment de l'entrée des utilisateurs dans ces cellules. Ce qui nécessite de savoir à priori (1) les chemins que les utilisateurs emprunteront et ; (2) les temps de transferts intercellulaires. Alors, afin d'atteindre notre objectif qui est de supporter les transferts intercellulaires sans l'arrêt forcé des sessions en cours dans les WCNs, nous avons proposé, dans la deuxième contribution, un modèle de prédiction individuelle de la mobilité des utilisateurs. Dans la troisième contribution, nous

avons proposé un modèle de prédiction collective/globale de la mobilité des utilisateurs. La troisième contribution est une solution au problème d'évolutivité de la deuxième contribution. Cependant, la deuxième contribution offre une meilleure précision sur la prédiction de la mobilité des utilisateurs ; donc une meilleure estimation de la bande passante disponible dans les cellules.

Dans la deuxième contribution, présentée au chapitre 4, nous avons proposé (1) un modèle individuel d'estimation des temps de transferts intercellulaires ; (2) un modèle d'estimation des quantités de bande passante disponible dans les cellules et ; (3) un système de contrôle d'admission des sessions d'applications ou services. Dans cette contribution, nous supposons, à priori, la connaissance des chemins (séquence de segments de route) qui seront utilisés par les utilisateurs grâce à la première contribution. Nous supposons aussi disposer du temps nécessaire à une session pour être complétée. En nous basant sur la densité de la zone de déplacement/navigation, les caractéristiques des mouvements des utilisateurs sur les segments de route, le planning d'apparition des couleurs des feux tricolores et les délais d'arrêt des utilisateurs aux panneaux d'arrêt obligatoires, nous avons estimé, pour chaque point de transfert intercellulaire sur le chemin d'utilisateur, un intervalle de temps pendant lequel celui-ci pourrait effectuer son transfert intercellulaire. Puis, sachant les sessions en cours (enregistrés préalablement par le réseau), les durées restantes avant d'être complétées des sessions en cours et les intervalles de temps de transferts intercellulaires des utilisateurs, nous avons fourni une estimation de la quantité de bande passante qui pourrait être disponible dans chaque cellule du réseau à un temps précis à l'avance. En d'autres termes, nous avons défini une fonction d'estimation de la bande passante qui pourrait être disponible dans une cellule en fonction d'un temps pris dans le futur. Enfin, en utilisant cette fonction, nous avons proposé un système distribué de CAC, à la fois orienté-mobilité et orienté-cellule. L'idée de l'utilisation d'une approche orientée-cellule dans notre contribution, dont l'approche principale est l'approche orientée-mobilité, est de réduire le taux de refus des demandes de nouvelles sessions dû aux erreurs de prédiction de la mobilité. Les résultats de simulation ont montré que notre contribution réduit conséquemment le taux (quasi nul) d'arrêt forcé et inopiné des sessions des applications et services en cours. Ils ont également montré que le taux de refus de demandes de nouvelles sessions et le taux d'utilisation de la bande passante sont

sensiblement égaux à ceux des meilleures contributions trouvées dans la littérature. Ainsi, notre contribution répond aux exigences d'un CAC efficace, supportant la QdS aussi bien pour les utilisateurs que pour les fournisseurs de WCNs.

Dans la troisième contribution, présentée au chapitre 5, nous avons proposé une plateforme d'intégration de modèles de prédiction collectives de la mobilité des utilisateurs (les chemins qui seront utilisés et les temps de transferts intercellulaires) avec un modèle d'estimation des quantités de bande passante disponible dans les cellules. Plus spécifiquement, après avoir identifié les sous zones à forte densité d'utilisateurs, nous avons déterminé des chemins entre elles selon les profils de route. Chaque chemin est obtenu en évaluant la possibilité pour les utilisateurs d'un même profil de route d'utiliser un segment de route. Cette évaluation est basée sur une extension du *second ordre de Markov* et utilise les données historiques sur les déplacements des utilisateurs. Ainsi, sachant le profil d'un utilisateur (calculé en utilisant les données historiques sur ses déplacements), nous sélectionnons son futur chemin selon sa destination (obtenue grâce à la première contribution). Pour l'estimation des temps de transferts intercellulaires, nous avons bâti une table des temps de transit pour chacun des segments de route de la zone de navigation. Pour le faire, nous avons divisé chaque segment de route en plusieurs sous-segments. Ensuite, en nous basant sur les caractéristiques des mouvements (vitesses courantes et temps de transit des sous-segments précédents) des utilisateurs présents dans les sous-segments, nous avons calculé la distribution de probabilité du temps qu'il faudrait à un utilisateur pour atteindre, à partir d'un sous-segment de route, chaque segment de route adjacente. La fonction cumulative de probabilité obtenue à partir de cette distribution de probabilité nous a permis d'estimer le temps de transit d'un sous-segment de route à chaque segment de route adjacente. Les temps de transit des sous-segments d'un segment de route forment une table des temps. Ainsi, à partir des tables des temps de transit des segments de route, nous avons estimé le temps de transit entre deux positions quelconques dans la zone de navigation. Puis, comme dans la deuxième contribution, nous avons fourni une estimation de la quantité de bande passante qui pourrait être disponible dans chaque cellule à un temps précis à l'avance. Enfin, sachant les quantités de bande passante disponible dans les cellules à l'avance, une demande de nouvelle session sera acceptée s'il y a suffisamment de

bande passante pour satisfaire cette demande le long du chemin du demandeur ; sinon, la demande sera refusée.

6.2. Perspectives et travaux futurs

Les contributions proposées dans cette thèse ouvrent plusieurs pistes de recherche pour des travaux futurs.

Premièrement, pour les simulations des contributions proposées, les modèles utilisent des données historiques de déplacements qui ne correspondent pas vraiment aux types de données nécessaires pour conduire des simulations convenables. Par exemple, pour la prédiction de la mobilité, nous avons utilisé des traces historiques des cellules visitées par les utilisateurs au lieu des segments de route. Aussi, pour la prédiction des temps de transferts intercellulaires, nous avons utilisé des traces générées par des algorithmes. Quand bien même ces traces sont proches de la réalité, elles ne peuvent pas représenter la situation réelle du comportement des automobilistes sur les routes et prendre en compte les incidents qui peuvent se produire. Alors, il serait souhaitable de collecter nos données historiques afin d'effectuer les simulations qui respectent les exigences des modèles proposés.

Deuxièmement, au-delà des simulations faites en laboratoire, nous envisageons des simulations sur des WCNs grandeur nature. Cela nous permettra de mieux évaluer les performances de nos modèles de prédiction. Cela nous permettra également de confirmer ou non que nos contributions produisent de bon résultats quel que soit l'environnement des simulations. Effectivement, les simulations faites en laboratoire ne sont pas suffisantes pour convaincre la communauté de la recherche de l'amélioration des performances que nos contributions apportent aux supports de la QdS dans les WCNs.

Troisièmement, nous prévoyons de travailler sur une mise en œuvre réelle de notre approche prédictive de la réservation de bande passante dans les réseaux 3GPP. Plus précisément, nous envisageons intégrer notre approche avec une fonction de détection des points d'accès afin de proposer un outil similaire au ANDSF (Access Network Node Discovery Function) [131-133]. En effet, notre proposition de ANDSF pourra être utilisée pour (a) collecter les caractéristiques de mobilité des utilisateurs grâce à la prédiction des

temps de transferts intercellulaires ; (b) prévoir la bande passante disponible le long des trajectoires prévues des utilisateurs ; et (c) recommander aux utilisateurs mobiles, le taux le plus approprié pour recevoir une session IP ou tout simplement de rejeter la demande de la session IP.

Toujours dans l'optique d'une mise en œuvre réelle de notre travail dans les réseaux 3GPP, nous prévoyons d'intégrer notre modèle de prédiction de la mobilité à l'orientation des antennes des stations de bases. Il faut noter qu'une station de base possède en général trois antennes. Chacune des antennes émet dans un angle de 120° ; soit 360° au total pour les trois antennes ensemble afin de couvrir une zone circulaire. Chaque antenne supporte donc un tiers des communications. Alors, il peut arriver qu'une antenne soit moins sollicitée tandis que les deux autres n'arrivent pas à satisfaire les utilisateurs se trouvant dans leur zone de couverture radio. En utilisant notre modèle de prédiction des chemins des utilisateurs, nous pourrions proposer un modèle d'orientation des antennes des stations de base en fonction des jours de la semaine et des moments de la journée.

Enfin, nous envisageons introduire nos modèles de prédiction de la mobilité aux réseaux véhiculaires VANETs [7]. L'objectif des VANETs est de contribuer à des routes plus sûres et plus efficaces à l'avenir, en fournissant des informations opportunes aux automobilistes et aux autorités intéressées. Dans les VANETs, les véhicules communiquent entre eux par l'intermédiaire de la communication inter-véhiculaire. Les véhicules communiquent aussi avec les équipements de la route par l'intermédiaire de la communication d'équipement-à-véhicule. L'un des problèmes dans les VANETS, c'est que dans le cadre de la communication d'équipement-à-véhicule, il faut trouver le véhicule le mieux localisé par rapport aux autres véhicules pour jouer le rôle d'intermédiaire (*Passerelle*). Ce véhicule est responsable de l'acheminement des données des autres véhicules vers les équipements. Alors, prédire d'avance la localisation des véhicules pourrait permettre de choisir les passerelles bien avant et de changer de passerelles en fonction du temps et des conditions du trafic routier. L'intégration ne notre modèle de la prédiction de la mobilité dans les VANETs contribuera énormément à l'amélioration de leur performance.

Bibliographie

[1] A. Damnjanovic, *et al.*, "A survey on 3GPP heterogeneous networks," *in IEEE Wireless Communications,* Vol. 18, No.3, pp. 10-21, Jun. 2011.

[2] A. Ghosh, *et al.*, "LTE-advanced: next-generation wireless broadband technology," *in IEEE Wireless Communications,* Vol. 17, No. 3, pp. 10-22, Jun. 2010.

[3] Alcatel Lucent, "The LTE Network Architecture—A Comprehensive Tutorial," 2009.

[4] J. Yick, B. Mukherjee and D. Ghosal, "Wireless sensor network survey," *in Computer Networks,* Vol. 52, No. 12, pp. 2292-2330, Aug. 2008.

[5] I. F. Akyildiz, X. Wang and W. Wang, "Wireless mesh networks: a survey," *in Computer Networks,* Vol. 47, No. 4, pp. 445-487, Mar. 2005.

[6] M. Boushaba, A. Hafid and A. Benslimane, "High accuracy localization method using AoA in sensor networks," *in Computer Networks,* Vol. 53, No. 18, pp. 3076-3088, Dec. 2009.

[7] H. Hartenstein and K. P. Laberteaux, "A tutorial survey on vehicular ad hoc networks," *in IEEE Communications Magazine,* Vol. 46, No. 6, pp. 164-171, Jun. 2008.

[8] V. Devarapalli, *et al.*, "RFC 3963 : Network Mobility (NEMO) Basic Support Protocol," *(Proposed Standard)* http://tools.ietf.org/pdf/rfc3963.pdf, Jan. 2005.

[9] A. Sayenko, *et al.*, "Ensuring the QoS requirements in 802.16 scheduling," in *Proc. of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems*, Terromolinos, Spain, Oct. 2006.

[10] J. Silvestre-Blanes, *et al.*, "Online QoS Management for Multimedia Real-Time Transmission in Industrial Networks," *in IEEE Transactions on Industrial Electronics,* Vol. 58, No. 3, pp. 1061-1071, Mar. 2011.

[11] P. McGovern, *et al.*, "Endpoint-Based Call Admission Control and Resource Management for VoWLAN," *in IEEE Trans. on Mobile Computing,* Vol. 10, No. 5, pp. 684-699, May 2011.

[12] P. Bellavista, *et al.*, "Self-adaptive handoff management for mobile streaming continuity," *in IEEE Transactions on Network and Service Management,* Vol. 6, No. 2, pp. 80-94, Jun. 2009.

[13] F. De Rango, P. Fazio and S. Marano, "Utility-Based Predictive Services for Adaptive Wireless Networks With Mobile Hosts," in *IEEE Transactions on Vehicular Technology,* Vol. 58, No. 3, pp. 1415-1428, Mar. 2009.

[14] J. D. Mallapur, *et al.*, "Fuzzy Based Bandwidth Management for Wireless Multimedia Networks Information Processing and Management." Vol. 70, V. V. Das, *et al.*, Eds.: Springer Berlin Heidelberg, Apr. 2010, pp. 81-90.

[15] Union internationale des télécommunications, "La Mésure de la société de l'information: résumé analytique," 2013.

[16] A. Vassilya and A. Isik, "Predictive mobile-oriented channel reservation schemes in wireless cellular networks," *in Wirel. Netw.,* Vol. 17, No. 1, pp. 149-166, Jan. 2011.

[17] U. Javed, *et al.*, "Predicting handoffs in 3G networks," *in SIGOPS Oper. Syst. Rev.,* Vol. 45, No. 3, pp. 65-70, Dec. 2011.

[18] K. P. Meetei, *et al.*, "Design and Development of a Handoff Management System in LTE Networks using Predictive Modelling," *in SASTECH,* Vol. 8, No. 2, Sep. 2009.

[19] Z. Becvar, P. Mach and B. Simak, "Improvement of handover prediction in mobile WiMAX by using two thresholds," *inComputer Networks,* Vol. 55, No. 16, pp. 3759-3773, Nov. 2011.

[20] H. Jeung*, et al.*, "Path prediction and predictive range querying in road network databases," in *The VLDB Journal,* Vol. 19, No. 4, pp. 585-602, May 2010.

[21] A. Nadembega, A. Hafid, and T. Taleb, "A Path Prediction Model to Support Mobile Multimedia Streaming," in *Proc. IEEE ICC*, Ottawa, Ontario, CANADA, Jun. 2012.

[22] M. Boc, M. Dias de Amorim and A. Fladenmuller, "Near-zero triangular location through time-slotted mobility prediction," in*Wireless Networks J.,* Vol. 17, No. 2, pp. 465-478, Feb. 2011.

[23] O. Mazhelis, "Real-time recognition of personal routes using instance-based learning," in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, Baden-Baden, Baden-Württemberg, Germany, Jun. 2011.

[24] J. Byungjin*, et al.*, "A Smart Handover Decision Algorithm Using Location Prediction for Hierarchical Macro/Femto-Cell Networks," *in Proc. IEEE VTC-Fall*, San Francisco, CA, USA, Sep. 2011.

[25] Xie Haitao and M. Xiangwu, "User Mobility Prediction Method Based on Spatial Cognition and ContextAwareness," in *J. of Convergence Information Tech.,* Vol. 6, No. 10, pp. 347-354, Oct. 2011.

[26] Q. Lu and P. Koutsakis, "Adaptive Bandwidth Reservation and Scheduling for Efficient Wireless Telemedicine Traffic Transmission," *in IEEE Trans. on Vehicular Technology,* Vol. 60, No. 2, pp. 632-643, Feb. 2011.

[27] K. Madhavi, R. K. Sandhya and R. P. Chandrasekhar, "Optimal Channel Allocation Algorithm with Efficient Bandwidth Reservation for Cellular Networks," *in Int'l J.of Computer Applications,* Vol. 25, No. 5, pp. 40-44, Jul. 2011.

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

[29] N. Samaan and A. Karmouch, "A mobility prediction architecture based on contextual knowledge and spatial conceptual maps," *in IEEE Trans. on Mobile Computing,* Vol. 4, No. 6, pp. 537-551, Nov.-Dec. 2005.

[30] A. Nadembega, A. Hafid, and T. Taleb, "A Destination Prediction Model based on Historical Data, Contextual Knowledge and Spatial Conceptual Maps," in *Proc. IEEE ICC*, Ottawa, Ontario, CANADA, Jun. 2012.

[31] A. Nadembega, A. Hafid, and T. Taleb, "Handoff Time Estimation Model for Vehicular Communications," in *Proc. IEEE ICC*, Budapest , Hungary, Jun. 2013.

[32] S. Deering and R. Hinden, "RFC 2460 : Internet Protocol, Version 6 (IPv6)," *(Proposed Standard)* http://tools.ietf.org/pdf/rfc2460.pdf, Dec. 1998.

[33] D. Johnson, C. Perkins and J. Arkko, "RFC 3775 : Mobility Support in IPv6," *(Proposed Standard)* http://tools.ietf.org/pdf/rfc3775.pdf, Jun. 2004.

[34] H. Soliman*, et al.*, "RFC 5380 : Hierarchical Mobile IPv6 (HMIPv6) Mobility Management," *(Proposed Standard)* http://tools.ietf.org/pdf/rfc5380.pdf, Oct. 2008.

[35] R. Wakikawa*, et al.*, "RFC 5648 : Multiple Care-of Addresses Registration," *(Proposed Standard)* http://tools.ietf.org/pdf/rfc5648.pdf, Oct. 2009.

[36] J. Arkko, C. Vogt and W. Haddad, "RFC 4866 : Enhanced Route Optimization for Mobile IPv6," *(Proposed Standard)* *http://tools.ietf.org/pdf/rfc4866.pdf*, May 2007.

[37] C. Ng, *et al.*, "RFC 4889 : Network Mobility Route Optimization Solution Space Analysis," *(Proposed Standard)* *http://tools.ietf.org/pdf/rfc4889.pdf*, Jul. 2007.

[38] W. Eddy, W. Ivancic and T. Davis, "RFC 5522 : Network Mobility Route Optimization Requirements for Operational Use in Aeronautics and Space Exploration Mobile Networks," *(Proposed Standard)* *http://tools.ietf.org/pdf/rfc5522.pdf*, Oct. 2009.

[39] C. Ng, *et al.*, "RFC 4888 : Network Mobility Route Optimization Problem Statement," *(Proposed Standard)* *http://tools.ietf.org/pdf/rfc4888.pdf*, Jul. 2007.

[40] S. Gundavelli, *et al.*, "RFC 5213 : Proxy Mobile IPv6," *(Proposed Standard)* *http://tools.ietf.org/pdf/rfc5213.pdf* Aug. 2008.

[41] E. Nordmark and M. Bagnulo, "Shim6: Level 3 Multihoming Shim Protocol for IPv6," *Internet draft, draft-ietf-shim6-proto-12.txt, (work in progress),* *http://tools.ietf.org/pdf/draft-ietf-shim6-proto-12.pdf* Feb. 2009.

[42] H. Soliman, *et al.*, "Flow Bindings in Mobile IPv6 and NEMO Basic Support," *draft-ietf-mext-flow-binding-03.txt (work in progress)* *http://tools.ietf.org/pdf/draft-ietf-mext-flow-binding-06.pdf*, Jul. 2009.

[43] S. Hussain, Z. Hamid and N. S. Khattak, "Mobility management challenges and issues in 4G heterogeneous networks," in *Proc. of the first international conference on Integrated internet ad hoc and sensor networks*, Nice, France, May 2006.

[44] C. Tzung-Shi, C. Yen-Ssu and C. Tzung-Cheng, "Mining User Movement Behavior Patterns in a Mobile Service Environment,"*in IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans,* Vol. 42, No. 1, pp. 87-101, Jan. 2012.

[45] Q. Zheng, *et al.*, "Agenda driven mobility modelling," in *Int'l J. of Ad Hoc and Ubiquitous Computing,* Vol. 5, No. 1, pp. 22-36, Jan. 2010.

[46] D. Karamshuk, *et al.*, "Human mobility models for opportunistic networks," in *IEEE Commun. Mag.,* Vol. 49, No. 12, pp. 157-165, Dec. 2011.

[47] T. Anagnostopoulos, C. Anagnostopoulos and S. Hadjiefthymiades, "Efficient Location Prediction in Mobile Cellular Networks," *inInt'l J. of Wireless Information Networks,* Vol. 19, No. 2, pp. 97-111, Nov. 2012.

[48] H. Wei-Jen, *et al.*, "Modeling Spatial and Temporal Dependencies of User Mobility in Wireless Mobile Networks," in *IEEE/ACM Trans. on Networking,* Vol. 17, No. 5, pp. 1564-1577, Oct. 2009..

[49] W. Wanalertlak, *et al.*, "Behavior-based mobility prediction for seamless handoffs in mobile wireless networks," in *Wirel. Netw.,* Vol. 17, No. 3, pp. 645-658, Apr. 2011.

[50] H. Abu-Ghazaleh and A. S. Alfa, "Application of Mobility Prediction in Wireless Networks Using Markov Renewal Theory," *in IEEE Trans. on Vehicular Technology,* Vol. 59, No. 2, pp. 788-802, Feb. 2010.

[51] J. Kim and A. Helmy, "Poster abstract: the challenges of accurate mobility prediction for ultra mobile users," *in SIGMOBILE Mob. Comput. Commun. Rev.,* Vol. 13, No. 3, pp. 58-61, Jul. 2009.

[52] S. Michaelis and C. Wietfeld, "Comparison of User Mobility Pattern Prediction Algorithms to increase Handover Trigger Accuracy," in *Proc. IEEE VTC*, Melbourne,Victoria, Australia, May 2006.

[53] I. Butun, *et al.*, "Impact of mobility prediction on the performance of Cognitive Radio networks," in Proc. *IEEE WTS*, Tampa, FL, USA, Apr. 2010.

[54] S. Akoush and A. Sameh, "Mobile user movement prediction using bayesian learning for neural networks," in *Proc. of the Int'l conf. on Wireless communications and mobile computing*, Honolulu, Hawaii, USA, Aug. 2007.

[55] M. Kim, D. Kotz and S. Kim, "Extracting a mobility model from real user traces," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[56] L. Liao, *et al.*, "Learning and inferring transportation routines," *in Artificial Intelligence,* Vol. 171, No. 5-6, pp. 311-331, Apr. 2007.

[57] T. Anagnostopoulos, *et al.*, "Predicting the location of mobile users: a machine learning approach," in *Proc. of the 2009 int'l conf. on Pervasive services*, London, United Kingdom, jul. 2009.

[58] J. Markoulidakis, *et al.*, "Mobility modeling in third-generation mobile telecommunications systems," *in IEEE Personal Communications,* Vol. 4, No. 4, pp. 41-56, Aug. 1997.

[59] F. Bai and A. Helmy, "A survey of mobility models," in *Wireless Ad Hoc and Sensor Networks, Kluwer Academic Publishers*, 2004.

[60] C. Song, *et al.*, "Modelling the scaling properties of human mobility," in *Nat. Phys.,* Vol. 6, No. 10, pp. 818-823, Sep. 2010.

[61] L. Kyunghan, *et al.*, "SLAW: A New Mobility Model for Human Walks," in *IEEE INFOCOM* , Rio de Janeiro, Brazil, Jun. 2009.

[62] A. Mei and J. Stefa, "SWIM: A Simple Model to Generate Small Mobile Worlds," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, Jun. 2009.

[63] C. Boldrini and A. Passarella, "HCMM: Modelling spatial and temporal properties of human mobility driven by users' social relationships," *in Computer Communications,* Vol. 33, No. 1, pp. 1056-1074, Jun. 2010.

[64] V. Borrel, *et al.*, "SIMPS: Using Sociology for Personal Mobility," *in IEEE/ACM Transactions on Networking,* Vol. 17, No. 3, pp. 831-842, Jun. 2009.

[65] Y. Shusen, *et al.*, "Using social network theory for modeling human mobility," *Network,* in *IEEE Network,* Vol. 24, No. 5, pp. 6-13, Sep.-Oct. 2010.

[66] D. Fischer, K. Herrmann and K. Rothermel, "GeSoMo - A general social mobility model for delay tolerant networks," in *IEEE MASS,* San Francisco, CA, USA, Nov. 2010.

[67] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," *in ACM SIGMOBILE Mobile Computing and Communications Review,* Vol. 11, No. 3, pp. 59 - 70, Jul. 2007.

[68] D. Yuan, F. Jialu and C. Jiming, "Experimental analysis of user mobility pattern in mobile social networks," in *Proc. IEEE WCNC*, Cancun, Quintana Roo, Mar. 2011.

[69] F. Ekman, *et al.*, "Working day movement model," in *Proc. of the 1st ACM SIGMOBILE workshop on Mobility models for Networking Research (MobilityModels'08)*, Hong Kong, Hong Kong, China, May 2008.

[70] A. Roy, J. Shin and N. Saxena, "Entropy-based location management in long-term evolution cellular systems," in *ET Communications,* Vol. 6, No. 2, pp. 138-146, Jul. 2012.

[71] T. Melodia, D. Pompili and I. F. Akyldiz, "Handling Mobility in Wireless Sensor and Actor Networks," *in IEEE Tran. on Mobile Computing,* Vol. 9, No. 2, pp. 160-173, Feb. 2010.

[72] Z. Zhong*, et al.*, "Scalable Localization with Mobility Prediction for Underwater Sensor Networks," *in IEEE Transactions on Mobile Computing,* Vol. 10, No. 3, pp. 335-348, Mar. 2011.

[73] I. F. Akyildiz, X. Jiang and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *in IEEE Wireless Communications,* Vol. 11, No. 4, pp. 16-28, Aug. 2004.

[74] N. Eagle*, et al.*, "Inferring Social Network Structure using Mobile Phone Data," *in the National Academy of Sciences (PNAS),* Vol. 106, No. 36, pp. 15274-15278, Sep. 2009.

[75] M. Anisetti*, et al.*, "Map-Based Location and Tracking in Multipath Outdoor Mobile Networks," in *IEEE Trans. on Wireless Commun.,* Vol. 10, No. 3, pp. 814-824, Mar. 2011.

[76] F. Zhu and J. McNair, "Multiservice vertical handoff decision algorithms," *in EURASIP J. Wirel. Commun. Netw.,* Vol. 2006, No. 2, pp. 52-52, Apr. 2006.

[77] P. S. Prasad and P. Agrawal, "Movement prediction in wireless networks using mobility traces," in *Proc. IEEE CCNC*, Las Vegas, NV, USA, Jan. 2010.

[78] S. Hongbo*, et al.*, "Mobility prediction in cellular network using hidden Markov model," in *Proc. IEEE CCNC*, Las Vegas, NV, USA, Jan. 2010.

[79] S. Bellahsene and L. Kloul, "A New Markov-Based Mobility Prediction Algorithm for Mobile Networks Computer Performance Engineering." Vol. 6342, A. Aldini*, et al.*, Eds.: Springer Berlin / Heidelberg, 2010, pp. 37-50.

[80] B. D. Ziebart*, et al.*, "Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior," in *UbiComp*, Seoul, Korea, Sep. 2008.

[81] A. Monreale*, et al.*, "WhereNext: a location predictor on trajectory pattern mining," *in Proc. ACM SIGKDD*, Paris, France, Jun-Jul 2009.

[82] S. Reddy*, et al.*, "Using mobile phones to determine transportation modes," in *ACM Trans. Sen. Netw.,* Vol. 6, No. 2, pp. 1-27, Feb. 2010.

[83] M. Afanasyev*, et al.*, "Usage Patterns in an Urban WiFi Network," *in IEEE/ACM Transactions on Networking,* Vol. 18, No. 5, pp. 1359-1372, Oct. 2010.

[84] C. Song*, et al.*, "Limits of Predictability in Human Mobility," *in Science,* Vol. 327, No. 5968, pp. 1018-1021, Feb. 2010.

[85] C. Francesco, G. D Lorenzo and C. Ratti, "Human Mobility Prediction based on Individual and Collective Geographical Preferences," in *Int'l IEEE Conf. on ITSC*, Madeira Island, Portugal, Sep. 2010.

[86] K. L. Dias*, et al.*, "Approaches to resource reservation for migrating real-time sessions in future mobile wireless networks," *in Wirel. Netw.,* Vol. 16, No. 1, pp. 39-56, Jan. 2010.

[87] P. Lytrivis*, et al.*, "An Advanced Cooperative Path Prediction Algorithm for Safety Applications in Vehicular Networks," *in IEEE Trans. on Intelligent Transportation Systems,* Vol. 12, No. 3, pp. 669-679, Sept. 2011.

[88] X. Zhong*, et al.*, "A New Adaptive Channel Reservation Scheme for Handoff Calls in Wireless Cellular Networks," *Lecture Notes in Computer Science,* Vol. 2345, pp. 672-684, 2002.

[89] S. Choi and K. G. Shin, "Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks," *in IEEE Trans. on Parallel and Distributed Systems,* Vol. 13, No.9, pp. 882-897, Sep. 2002.

[90] M. M. Islam and M. Murshed, "Parametric mobility support dynamic resource reservation and call admission control scheme for cellular multimedia communications," *in Computer Communications,* Vol. 30, No. 2, pp. 233-248, Jan. 2007.

[91] S. Rashad*, et al.*, "User mobility oriented predictive call admission control and resource reservation for next-generation mobile networks," *in Journal of Parallel and Distributed Computing,* Vol. 66, No. 7, pp. 971-988, Jul. 2006.

[92] L. Duan-Shin and H. Yun-Hsiang, "Bandwidth-reservation scheme based on road information for next-generation cellular networks," *in IEEE Trans. on Vehicular Technology,* Vol. 53, No. 1, pp. 243-252, Jan. 2004.

[93] C.-J. Huang*, et al.*, "An adaptive bandwidth reservation scheme for 4G cellular networks using flexible 2-tier cell structure," *in Expert Systems with Applications,* Vol. 37, No. 9, pp. 6414-6420, Sep. 2010.

[94] A. Esmailpour and N. Nasser, "Dynamic QoS-Based Bandwidth Allocation Framework for Broadband Wireless Networks," *in IEEE Trans. on Vehicular Technology,* Vol. 60, No. 6, pp. 2690-2700, Jul. 2011.

[95] L. Shufeng*, et al.*, "Overlap Area Assisted Call Admission Control Scheme for Communications System," *in IEEE Trans. on Aerospace and Electronic Systems,* Vol. 47, No. 4, pp. 2911-2920, Oct. 2011.

[96] 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'l Conf. on SEISCON*, Chennai, Tamil Nadu, India, Jul. 2011.

[97] J. S. Wu*, et al.*, "Admission Control for Multiservices Traffic in Hierarchical Mobile IPv6 Networks by Using Fuzzy Inference System," *in Journal of Computer Networks and Communications,* Vol. 2012, 2012.

[98] Y. Jun, S. S. Kanhere and M. Hassan, "Improving QoS in High-Speed Mobility Using Bandwidth Maps," *in IEEE Trans. on Mobile Computing*, Vol. 11, No. 4, pp. 603-617, Apr. 2012.

[99] 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.

[100] S. N. Ahmed and B. Ferri, "Prediction based bandwidth reservation," in *IEEE CDC*, Atlanta, GA, USA, Dec. 2010.

[101] L. Mokdad, M. Sene and A. Boukerche, "Call Admission Control Performance Analysis in Mobile Networks Using Stochastic Well-Formed Petri Nets," *in IEEE Trans. on Parallel and Distributed Systems,* Vol. 22, No. 8, pp. 1332-1341, Aug. 2011.

[102] G. I. Tsiropoulos*, et al.*, "Probabilistic framework and performance evaluation for prioritized call admission control in next generation networks," *in Computer Communications,* Vol. 34, No. 9, pp. 1045-1054, Jun. 2011.

[103] 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.
[104] 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'l Symp. on Computer Commun. Control and Automation*, Tainan, Taiwan, May 2010.
[105] M. Ravichandran, P. Sengottuvelan and A. Shanmugam, "An Approach for Admission Control and Bandwidth Allocation in Mobile Multimedia Network Using Fuzzy Logic," *in Int'l J. of Recent Trends in Engineering,* Vol. 1, No. 1, pp. 289-293, May 2009.
[106] S. Kim, "Cellular network bandwidth management scheme by using nash bargaining solution," *in IET Communications,* Vol. 5, No. 3, pp. 371-380, Feb. 2011.
[107] T. Taleb, A. Hafid, and A. Nadembega, "Mobility-Aware Streaming Rate Recommendation System," *in Proc. of IEEE GLOBECOM'11*, Houston, Texas, USA, Dec. 2011.
[108] C.-Y. Wang, H.-Y. Huang. and R.-H. Hwang, "Mobility management in ubiquitous environments," i*n Personal Ubiquitous Comput.,* Vol. 15, No. 3, pp. 235-251, Mar. 2011.
[109] U. Rathnayake, *et al.*, "EMUNE: Architecture for Mobile Data Transfer Scheduling with Network Availability Predictions," *in Mob. Netw. Appl.,* Vol. 17, No. 2, pp. 216-233, Apr. 2012.
[110] P. S. Prasad and P. Agrawal, "A generic framework for mobility prediction and resource utilization in wireless networks," in *Proc. COMSNETS*, Bangalore, India, Jan. 2010.
[111] H. Chenn-Jung, *et al.*, "A Probabilistic Mobility Prediction Based Resource Management Scheme for WiMAX Femtocells," in *Proc. ICMTMA*, Changsha City, China, Mar. 2010.
[112] F. Yu, *et al.*, "Performance enhancement of combining QoS provisioning and location management in wireless cellular networks," *in IEEE Transactions on Wireless Communications,* Vol. 4, no. 3, pp. 943-953, May 2005.
[113] I. F. Akyildiz and W. Wang, "The predictive user mobility profile framework for wireless multimedia networks," in *IEEE/ACM Trans. Netw.,* Vol. 12, No. 6, pp. 1021-1035, Dec. 2004.
[114] C.-F. Wu, *et al.*, "A novel call admission control policy using mobility prediction and throttle mechanism for supporting QoS in wireless cellular networks," in *J. Control Sci. Eng.,* Vol. 2011, pp. 21-31, Jan. 2011.
[115] P. Salvador and A. Nogueira, "Markov Modulated Bi-variate Gaussian Processes for Mobility Modeling and Location Prediction NETWORKING 2011." Vol. 6640, J. Domingo-Pascual, *et al.*, Eds.: Springer Berlin / Heidelberg, 2011, pp. 227-240.
[116] P. Bellavista, A. Corradi and L. Foschini, "IMS-Compliant management of vertical handoffs for mobile multimedia session continuity," in *IEEE Communications Magazine,* Vol. 48, No. 4, pp. 114-121, Apr. 2010.
[117] U. Rathnayake, M. Ott and A. Seneviratne, "Network availability prediction with hidden context," in *Performance Evaluation,* Vol. 68, No.9, pp. 916-926, Sep. 2011.

[118] M. Lee*, et al.*, "Predictive Mobility Support with Secure Context Management for Vehicular Users Network Control and Engineering for QoS, Security and Mobility, IV." Vol. 229, D. Gaïti, Ed.: Springer Boston, 2007, pp. 21-28.
[119] L. Yinan and C. Ing-Ray, "Design and Performance Analysis of Mobility Management Schemes Based on Pointer Forwarding for Wireless Mesh Networks," in *IEEE Transactions on Mobile Computing,* Vol. 10, No. 3, pp. 349-361, Mar. 2011.
[120] N. Bouabdallah, R. Langar and R. Boutaba, "Design and Analysis of Mobility-Aware Clustering Algorithms for Wireless Mesh Networks," in *IEEE/ACM Trans. on Networking,* Vol. 18, No. 6, pp. 1677-1690, Dec. 2010.
[121] W. Viriyasitavat, F. Bai and O. K. Tonguz, "Dynamics of Network Connectivity in Urban Vehicular Networks," in *IEEE Jour'l on Selected Areas in Communications,* Vol. 29, No. 3, pp. 515-533, Mar. 2011.
[122] A. Rodriguez-Carrion, C. Garcia-Rubio and C. Campo, "Performance Evaluation of LZ-Based Location Prediction Algorithms in Cellular Networks," in *IEEE Communications Letters,* Vol. 14, No. 8, pp. 707-709, Aug. 2010.
[123] J. J. Pan*, et al.*, "Tracking Mobile Users in Wireless Networks via Semi-Supervised Colocalization," *in IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 34, No. 3, pp. 587-600, Mar. 2012.
[124] Y. Zheng*, et al.*, "Understanding transportation modes based on GPS data for web applications," in *ACM Trans. Web,* Vol. 4, No 1, pp. 1-36, Jan. 2010.
[125] K. Keshav and P. Venkataram, "A Dynamic Bandwidth Allocation Scheme for Interactive Multimedia Applications over Cellular Networks," i*n Porc. ICN*, St. Maarten, The Netherlands Antilles, Jan. 2011.
[126] N. Vallina-Rodriguez and J. Crowcroft, "Energy Management Techniques in Modern Mobile Handsets," *IEEE Communications Surveys & Tutorials,* Vol. 15, No. 1, pp. 179-198, First Quarter 2013.
[127] J. Sorber*, et al.*, "Tula: Balancing Energy for Sensing and Communication in a Perpetual Mobile System," in *IEEE Trans. on Mobile Computing,* Vol. 12, No. 4, pp. 804-816, Apr. 2013.
[128] C. Shi*, et al.*, "Serendipity: enabling remote computing among intermittently connected mobile devices," *in Proc.* of the thirteenth ACM international symposium on Mobile Ad Hoc Networking and Computing, Hilton Head, South Carolina, USA, Jun. 2012.
[129] E. Stevens-Navarro, L. Yuxia, and V. W. S. Wong, "An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *in IEEE TVT,* Vol. 57, No. 2, pp. 1243-1254, Mar. 2008.
[130] R. Baumann, F. Legendre and P. Sommer, "Generic mobility simulation framework (GMSF)," in Proc. ACM SIGMOBILE, Hong Kong, Hong Kong, China, May 2008.
[131] 3GPP Specifications, "Architecture enhancements for non-3GPP accesses," *TS 23.402*.
[132] 3GPP Specifications, "Operator Policies for IP Interface Selection (OPIIS)," *TR 23.853*.
[133] 3GPP Specifications, "Data Identification in Access Network Discovery and Selection Function (ANDSF) (DIDA)," *TR 23.855*.
[134] T. Taleb, A. Ksentini, and F. Filali, "Wireless Connection Steering for Vehicles," *in Proc. IEEE Globecom*, Anaheim, USA, Dec. 2012.

[135] T. Taleb and A. Ksentini, "QoS/QoE predictions-based admission control for femto communications," *in Proc. IEEE ICC,* Ottawa, Canada, Jun. 2012.
[136] A. Nadembega, A. Hafid, and T. Taleb, "A Framework for Mobility Prediction and High Bandwidth Utilization to Support Mobile Multimedia Streaming," in *Proc. IEEE ANTS*, Chennai, India, Dec. 2013.
[137] K. Sungwook and P. K. Varshney, "An integrated adaptive bandwidth-management framework for QoS-sensitive multimedia cellular networks," in *IEEE TVT,* Vol. 53, No 3, pp. 835-846, May 2004.
[138] A. Sgora and D. Vergados, "Handoff prioritization and decision schemes in wireless cellular networks: a survey," *in IEEE Communications Surveys & Tutorials,* Vol. 11, No. 4, pp. 57-77, Fourth quarter 2009.