UNIVERSITÉ DE MONTRÉAL

INTÉGRATION ET GESTION DE MOBILITÉ DE BOUT EN BOUT DANS LES RÉSEAUX MOBILES DE PROCHAINE GÉNÉRATION

ABDELLATIF EZZOUHAIRI

DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION

DU DIPLÔME DE PHILOSOPHIAE DOCTOR

(GÉNIE INFORMATIQUE)

DÉCEMBRE 2009

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

INTÉGRATION ET GESTION DE MOBILITÉ DE BOUT EN BOUT DANS LES RÉSEAUX MOBILES DE PROCHAINE GÉNÉRATION

présentée par : EZZOUHAIRI Abdellatif

en vue de l'obtention du diplôme de : Philosophiae Doctor

a été dûment accepté par le jury d'examen constitué de :

Mme BOUCHENEB Hanifa, Doctorat, présidente

M. QUINTERO Alejandro, Doct., membre et directeur de recherche

M. PIERRE Samuel, Ph. D., membre et codirecteur de recherche

Mme NICOLESCU Gabriela, Doct., membre

M. AJIB Wessam, Ph. D., membre externe

DÉDICACE

À toute ma famille …

REMERCIEMENTS

Je tiens à exprimer mes sincères remerciements à mon Directeur de recherche, le professeur Alejandro Quintero, pour la qualité de son encadrement, ses suggestions, sa disponibilité et son soutien financier tout au long de la préparation de ma thèse.

Mes remerciements vont aussi à mon Codirecteur le professeur Samuel Pierre Directeur du LARIM. Je tiens à lui exprimer ma profonde reconnaissance pour son encadrement, son appui et son soutien financier durant tout mon séjour au LARIM.

Mes remerciements s'adressent aussi à Mme Hanifa Boucheneb, Gabriela Nicolescu et Wessam Ajib d'avoir accepté d'évaluer cette thèse malgré leurs multiples occupations et engagements.

Mes remerciements vont aussi à tous les membres du LARIM, pour leur soutien et pour l'ambiance de travail.

Un gros merci à ma très chère femme qui n'a jamais cessé de m'encourager et de m'offrir l'ambiance familiale propice pour mener à bien mon travail.

Mes sincères remerciements à mes chers enfants : Meriem, Marwa et Amin qui, malgré leur jeune âge, ont su vivre avec mes préoccupations de recherche et d'études.

Enfin, je tiens à exprimer ma gratitude à tous ceux qui ont participé de près ou de loin à l'accomplissement de cette thèse.

RÉSUMÉ

Pendant les dix dernières années, l'utilisation des systèmes de communication sans fil est devenue de plus en plus populaire tant chez les entreprises que chez les particuliers. Cette nouvelle tendance du marché est due, en grande partie, à la performance grandissante des réseaux mobiles qui concurrencent davantage les réseaux filaires en termes de bande passante, de coût et de couverture. Toutefois, cette catégorie de solutions sans fil est conçue pour des services spécifiques et utilise des technologies très variées. De plus, les usagers sont de plus en plus mobiles et requièrent des applications sensibles au délai (voix, multimédia, etc.).

Dans ce nouveau contexte de mobilité, la prochaine génération des réseaux sans fil (4G) s'annonce comme l'ultime solution visant à satisfaire les exigences des usagers tout en tirant profit de la complémentarité des services offerts par les systèmes mobiles existants. Pour ce faire, la principale vocation de la future génération (4G) consiste en l'intégration et la convergence des technologies sans fil existantes et celles à venir. Cette intégration passe obligatoirement par l'utilisation du protocole IP (Internet Protocol) qui permet de cacher l'hétérogénéité des systèmes intégrés puisqu'il demeure l'unique couche commune à toutes les plateformes mobiles.

Plusieurs solutions d'intégration ont été proposées dans la littérature. Celles-ci concernent des architectures d'intégration et des mécanismes de gestion de mobilité. Cependant, les approches proposées ne font pas l'unanimité et souffrent de plusieurs handicaps liés, en particulier, à l'interopérabilité et la garantie des relèves sans coupures.

Partant de ce constat, notre principal point d'intérêt dans cette thèse consiste à proposer des solutions liées aux problématiques de conception et de déploiement des réseaux 4G. Plus

spécifiquement, nous nous concentrerons sur l'intégration et la gestion de mobilité dans un environnement mobile hétérogène.

Pour ce faire, cette thèse débute par une revue de littérature approfondie afin de déceler les limites des solutions existantes et de dresser de nouvelles pistes pour notre recherche. Ensuite, trois articles aborderont les principales problématiques identifiées dans le cadre de ce travail à savoir: la mobilité, la préparation des relèves et l'intégration. Plus précisément, le premier article propose un mécanisme de gestion de mobilité de bout en bout désigné par : Adaptive end-to-end mobility scheme for seamless horizontal and vertical handoffs. Le protocole proposé garantit des relèves sans coupure tout en améliorant le flux de données reçu (*throughput*) pendant la période de changement du point/réseau d'attache.

Le second article introduit une stratégie de décision de relève visant à assurer une bonne préparation du processus de relève avant de la déclencher. Concrètement, cette stratégie inclut un mécanisme d'analyse de contexte qui s'adapte aux exigences des environnements sans fil multi-accès. En outre, elle incorpore un processus d'initiation des relèves basé sur la logique floue qui permet de décider du moment et des conditions opportuns pour déclencher une relève. Par ailleurs, afin de garantir un meilleur choix du réseau de destination, ladite stratégie incorpore une fonction de préférence conçue spécialement pour que les destinations choisies puissent satisfaire les requis des usagers mobiles en termes de QdS et de stabilité.

Enfin, le troisième article propose une architecture hybride et interopérable qui permet d'intégrer différentes technologies mobiles autour d'une dorsale IP. Cette architecture s'avère bien adaptée aux exigences des réseaux métropolitains de la prochaine génération dans la mesure où elle est évolutive, économique et garantit la connexion au meilleur réseau disponible lors d'une mobilité horizontale ou verticale. De plus, l'itinérance globale des usagers est rendue plus accessible moyennant des accords de services établis directement avec une tierce autorité au lieu

des accords bilatéraux conventionnels. En guise de récapitulation, les principales contributions de cette thèse se résument comme suit:

- proposition d'un mécanisme de gestion de mobilité de bout en bout qui tient compte de la mobilité locale et globale au niveau transport et qui vise à réduire le délais des relèves, la perte des paquets et la charge de signalisation sur le réseau. De plus, le problème de détérioration du flux de données reçues après l'exécution d'une relève a été traité.
- proposition d'une stratégie de préparation de relève basée sur la logique floue, le but étant de déterminer les conditions opportunes pour identifier et initier aussi bien des relèves forcées que volontaires. De plus, notre solution permet de choisir des destinations appropriées et incorpore une architecture d'analyse de contexte qui garantit la disponibilité et la confidentialité des informations échangées à travers des systèmes et des environnements hétérogènes.
- développement d'une nouvelle fonction de préférence qui considère un nombre variable de paramètres de contexte et qui tient compte également de la stabilité des réseaux lors du choix d'une destination.
- conception d'une architecture d'intégration interopérable pour les réseaux métropolitains. Cette architecture est ouverte et peut supporter aussi bien la mobilité au niveau IP qu'au niveau transport.
- proposition d'une version améliorée du protocole HTM de manière à garantir la qualité de service en incluant les phases de préparation des relèves et du choix des réseaux de destination.
- validation des solutions proposées moyennant des simulations et des modèles théoriques.

ABSTRACT

During the last few years, the use of wireless systems is becoming more and more popular. This tendency can be explained by the fact that mobile technologies are gaining in performance in terms of bandwidth, coverage and cost compared to the traditional wired solutions. However, each mobile network is tailored for a specific type of services and users. Moreover, end users are expected to become more and more mobile and show an increasing interest to real-time applications. In these circumstances, the next generation of mobile networks (4G) appears to be the ultimate solution that will satisfy mobile user demands and take benefit of the existing wireless systems. Indeed, the future generation consists of integrating, in an intelligent manner, the existing/future wireless systems in a way that users can obtain their services via the best available network.

This integration passes through the use of the Internet Protocol (*IP*) that will hide the heterogeneity pertaining to the integrated networks. To deal with this very important task, several solutions are available in the literature. The proposed approaches cover some basic topics such as interworking architecture and mobility management. Nevertheless, these proposals suffer from drawbacks relevant to the guarantee of QoS through heterogeneous technologies.

Based on these facts, the main concern of this thesis is to propose new solutions that address some well known problems pertaining to the conception and deployment of the next generation of mobile networks. More specifically, this work starts with a deep analysis of the existing solutions in order to identify new hints for our research. Then, we present three articles which refer to the main themes concerned by this thesis. Particularly, the first article proposes an efficient end-to-end mobility management scheme called: Adaptive end-to-end mobility scheme

for seamless horizontal and vertical handoffs. The proposed protocol ensures seamless handovers and improves the throughput during the handoff period.

The second article introduces a new Handoff Decision Strategy designated as HDS. This strategy is based on the Fuzzy logic and includes an efficient context analysis scheme that guarantees the availability and the privacy of context parameters through heterogeneous mobile systems. Furthermore, HDS implements a handoff initiation process that allows a mobile user to decide which type of handoff to trigger (i,e., forced or voluntary) and under which conditions. In addition, HDS incorporates a powerful network selection mechanism that allows a mobile node to be always connected to the best available network when it performs either inter-system or intra-system handoffs.

Finally, the third article proposes a new Hybrid Interworking Architecture for metropolitan networks (HIA). The main objective of HIA consists of integrating any type of existing/future mobile systems while hiding their heterogeneity from each other. Additionally, HIA guarantees inter-system authentication and billing by using an independent authority referred to as Interworking Cooperation Server (ICS). Furthermore, to ensure seamless global roaming, HIA is coupled with an efficient transport layer mobility scheme that uses underlying hints to prepare appropriate handoffs. Performance analysis show that the proposed architecture respects well the 4G requirements in terms of cost, deployment and global roaming compared to the existing solutions. As a conclusion, the main contributions of this thesis consist of:

- proposing an end-to-end mobility scheme that takes into account micro and macro mobility at the transport level. This proposal aims to reduce handoff delay, packet loss, and signaling load. Moreover, it addresses the problem of deterioration of throughput due to spurious retransmission during handoff periods.

- introducing a handoff decision strategy which uses Fuzzy Logic to prepare and initiate appropriate handoffs. Furthermore, this strategy integrates an efficient context-aware architecture that guarantees context information privacy through heterogeneous mobile systems.
- developing a new preference function that considers a wide range of context parameters and takes into account network stability.
- conceiving an interworking architecture that integrates metropolitan networks with respect to 4G requirements in terms of mobility and service continuity.
- proposing an enhanced Hierarchical Transport layer Mobility (HTM) scheme that guarantees QoS requirements by incorporating handoff preparation and network selection into the handoff process.
- evaluating the effectiveness of the proposed solutions through simulations and theoretical models.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

3G	:	Third Generation
3GAAA	:	Third Generation AAA
3GPP	:	Third Generation Partnership Project
3GPP2	:	Third Generation Partnership Project 2
4G	:	Fourth Generation
AAA	:	Authentication Authorization and Accounting
ABC	:	Always Best Connected
AccM	:	Accounting Module
AHP	:	Analytic Hierarchy Process
AM	:	Authentication Module
AMU	:	Anchor Mobility Unit
AP	:	Access Point
AR	:	Access Router
ARR	:	Account Register Record
AuM	:	Authentication Module
BER	:	Bit Error Rate
BLR	:	Boundary Location Register
BR	:	Boreder Router
BSC	:	Base Station Controller
BTS	:	Base Tranceiver Station

BU	:	Binding Update
CARD	:	Candidate Access Router Discovery
CAS	:	Context Aware Server
CBR	:	Constant Bit Rate
CM	:	Cooperation Module
CN	:	Correspondent Node
CoM	:	Cooperation Module
CWND	:	Congestion Window
DCF	:	Distributed Coordination Function
DDNS	:	Dynamic DNS
DNS	:	Domain Name Sevice
E-PCF	:	Extended PCF
E-SGSN	:	Extended SGSN
F-HMIPv6	:	Fast Handoff for Hierarchical Mobile IPv6
FL	:	Fuzzy Logic
FMIPv6	:	Fast handovers for Mobile IPv6
FTP	:	File Transfert Protocol
FY	:	Forced Yes
GGSN	:	Gateway GPRS Support Node
GIF	:	GPRS Interworking Function
GPRS	:	General Packet Radio Service
GRA	:	Grey Relational Analysis
GRX	:	GPRS Roaming eXchange

GSM	:	Global System for Mobile communications
HA	:	Home Agent
HDLM	:	Hierarchical DDNS Location Management
HDS	:	Handoff Decision Strategy
HIA	:	Hybrid Interworking Architecture
HiperLAN	:	High Performance Radio LAN
HMIPv6	:	Hierarchical Mobile IPv6
HN	:	Home Network
HTM	:	Hierarchical Transport layer Mobility
IA	:	Information Analyzer
ICS	:	Interworking Cooperation Server
IEEE	:	Institute of Electrical and Electronics Engineers
IETF	:	Internet Engineering Task Force
IG	:	Interworking Gateway
IGN	:	Interworking Gateway Node
IM	:	Information Manager
IP	:	Internet Protocol
LAAA	:	Local Authentication Authorization and Accounting
LAN	:	Local Area Network
LCoA	:	on-Link Care-of-Address
LICS	:	Local Interworking Cooperation Server
LUP	:	Local User Profile
MADM	:	Multiple Attribute Decision Making

MANET	:	Mobile Ad hoc NETwork
MAP	:	Mobility Anchor Point
MIP (v4 / v6)	:	Mobile IP (version 4/6)
MN	:	Mobile Node
mSCTP	:	Mobile Stream Control Transmission Protocol
NAR	:	New Access Router
NDS_Req	:	New Destination Selection Request
NGMN	:	Next Generation Mobile Network
NIA	:	Network Interoperating Agent
NN	:	Neural Network
Ns-2	:	Network Simulator version 2
NSA	:	Network Security Agreement
NsaM	:	NSA Module
PAR	:	Previous Access Router
PCF	:	Packet Control Function
PDA	:	Personnal Data Assistant
PDSN	:	Packet Data Serving Node
QdS (QoS)	:	Qualité de Service (Quality of Service)
RCoA	:	Regional Care-of-Address
RFC	:	Request For Comment
RNC	:	Radio Network Controller
RSS	:	Received Signal Strength
RTO	:	Retransmission Time Out

SACK	:	Selective Acknowledgment
SCTP	:	Stream Control Transmission Protocol
SGSN	:	Serving GPRS Support Node
SIP	:	Session Initiation Protocol
SIR	:	Signal to Interface Ratio
SLA	:	Service Level Agreement
SMD	:	Signal Measurement Device
SMR	:	Session to Mobility Ratio
SS	:	Storage Support
SSTHRESH	:	Slow Start Threshold
TCP	:	Transmission Control Protocol
TOPSIS	:	Technique for Ordering Preference by Similiraty Ideal Solution
UDP	:	User Datagram Protocol
UMTS	:	Universal Mobile Telecommunication Systems
VY	:	Voluntary Yes
WAF	:	WLAN Adaptation Function
WG	:	Wireless Gateway
WiFi	:	Wireless Fidelity
WiMax	:	Worldwide Interoperability for Microwave Access
WLAN	:	Wireless Local Area Network

CHAPITRE 1 INTRODUCTION

Les systèmes de communication sans fil ont, récemment, connu d'importants progrès aussi bien sur le plan des infrastructures que sur celui des services offerts aux usagers. L'entrée en marché des réseaux mobiles de troisième génération (3G) et l'apparition d'équipements mobiles hyper sophistiqués (PDA : Personal Data Assistant, ordinateurs portatifs, iphones, etc.), ainsi que l'émergence des réseaux sans fil tels que: 802.11x, Bluetooth, HyperLan, WiMax, etc., reflète bien cette nouvelle tendance du marché. Chacun de ces réseaux est conçu pour une catégorie particulière d'usagers et de services. Dès lors, notre paysage de communication est devenu de plus en plus hétérogène tout en offrant une visible complémentarité en termes de services, de couverture et de coût. Afin de tirer profit de cette diversité, de nouvelles approches d'interopérabilité s'avèrent nécessaires. C'est dans ce sens qu'on a initié la recherche dans un nouveau concept de réseaux mobiles désigné par quatrième génération. La principale vocation de cette nouvelle génération consiste en la convergence et l'intégration des différentes plates-formes mobiles autour d'une dorsale IP (*Internet Protocol*). L'utilisation du protocole IP permet de cacher l'hétérogénéité des systèmes intégrés puisque celui-ci demeure l'unique couche commune aux différentes technologies sans fil. De plus, la tendance actuelle dans les réseaux de communications privilégie la commutation des paquets lors de l'acheminement des données, ce qui favorise davantage l'adoption d'un support d'interopérabilité basé IP. Dans la pratique, cette intégration s'annonce difficile et fait face à de sérieux défis. Parmi ceux-ci, on cite la conception de nouveaux terminaux mobiles susceptibles de supporter différentes technologies d'accès, la gestion de mobilité, la garantie de qualité de service à travers des environnements mobiles hétérogènes, la sécurité, la facturation, etc. Dans le cadre de cette thèse, nous nous focaliserons, en particulier, sur la gestion de mobilité et l'intégration des systèmes hétérogènes.

1.1 Définitions et concepts de base

L'enjeu majeur de la prochaine génération de réseaux mobiles, communément désignée par quatrième génération ou 4G, est d'offrir aux usagers mobiles de bonnes performances en termes de connectivité, de mobilité et de service. Plus précisément, les abonnés des réseaux mobiles 4G auront la possibilité d'engager des services et d'exécuter des applications demandant des exigences élevés de QdS. Autrement dit, les applications temps-réel prendront une place plus imposante dans la priorité des utilisateurs de la future génération.

Selon Akyildiz et al. (2005), deux principales stratégies peuvent être envisagées pour la conception de la prochaine génération de réseaux mobiles. La première consiste à développer un nouveau système sans fil incluant un réseau cœur et un réseau d'accès. Cette option est loin d'être réaliste puisqu'elle requière le remplacement des systèmes actuels. La deuxième possibilité, plus réaliste, vise à intégrer de manière intelligente les systèmes sans fil existants et ceux à venir de façon à ce que les usagers mobiles aient toujours accès au meilleur réseau disponible. En effet, dans ce dernier cas, les abonnés profiteront de la complémentarité des services offerts par chacun des réseaux intégrés. De plus, l'utilisation des infrastructures existantes permet de réduire considérablement les délais et les coûts de déploiement. Par ailleurs, l'usage universel du protocole IP renforce davantage l'approche d'intégration vu que celui-ci permettra l'interopérabilité des différents systèmes mobiles tout en cachant leur hétérogénéité.

Pratiquement, on s'attend à ce que la prochaine génération de réseaux mobiles puisse tenir compte des points suivants:

- garantir aux usagers une meilleure connectivité en tout temps (*always best connected*);

- utiliser des terminaux multi-interfaces ou multi-modes pour pouvoir se connecter à différentes technologies d'accès radio;
- être évolutive tout en maintenant en service les infrastructures existantes;
- assurer de bonnes conditions de sécurité;
- se déployer à coût minimal;
- supporter la mobilité à travers des technologies hétérogènes.

La Figure 1.1 est une illustration sommaire de la prochaine génération de réseaux mobiles.

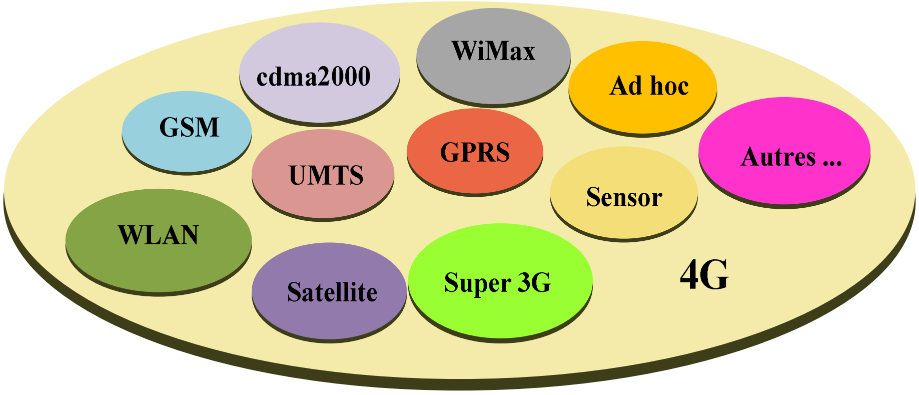


Figure 1.1 Exemple de réseaux 4G

Le concept de réseau basé IP ou tout IP fait référence à l'utilisation du protocole IP de bout-en-bout d'une chaîne de communication, sans que les données transmises aient à transiter par des réseaux utilisant d'autres protocoles (au niveau réseau). Ce concept est considéré comme l'élément fédérateur des réseaux mobiles de la 4G dans le sens où le protocole IP est indépendant de la technologie d'accès radio.

La mobilité réfère au fait de se procurer un service indépendamment de la localisation et du mouvement (Pierre, 2007). En d'autres termes, c'est la possibilité pour qu'un usager mobile puisse accéder à l'ensemble des services auxquels il est abonné sans se préoccuper de sa localisation ni de son mouvement.

Quant à la gestion de mobilité, elle peut être divisée en: gestion de localisation et gestion de relève. La gestion de localisation vise à identifier la position courante du mobile au sein de son réseau d'attache. Cette opération est assurée par l'échange périodique d'informations de localisation (*location update*) entre le mobile et le réseau.

La relève ou gestion de relève est un processus qui permet à un usager mobile de maintenir sa connexion active tout en changeant son canal de communication d'un point d'accès à un autre.

Dans le cadre des réseaux 4G, on s'attend à ce que la relève soit aussi bien forcée que volontaire. Une relève est dite forcée si le mobile est sérieusement menacé de perdre sa connexion suite à une détérioration de la puissance du signal reçu (RSS), une dégradation significative de la bande passante ou à tout autre raison. En revanche, une relève volontaire vise à améliorer la préférence de l'usager en termes de qualité de service (QdS) ou de coût même si le mobile bénéficie d'une connectivité satisfaisante dans son réseau d'attache.

Lorsqu'un mobile change de point d'attache à l'intérieur d'un même domaine administratif, on parle de relève intra-domaine ou micro-mobilité. Par contre, lorsque le mobile se déplace à travers différents domaines, on parle de relève inter-domaine ou macro-mobilité. Par ailleurs, due à la coexistence de différentes technologies sans fil dans les réseaux 4G, on distingue deux principales catégories de relèves désignées par: relève horizontale et relève verticale. Une relève est horizontale ou intra-système/intra-technologie si l'ancien et le nouveau point d'attache du mobile appartiennent à la même technologie. Autrement, la relève est considérée comme verticale ou inter-system/inter-technologie.

La procédure de relève peut être divisée en deux phases: préparation et exécution. La préparation d'une relève inclut l'initiation et la décision. Tandis que l'exécution réfère à l'utilisation de mécanismes de mobilité pour accomplir le processus de relève. Dans les réseaux

4G, l'initiation d'une relève ne sera plus conditionnée uniquement par la qualité du signal reçu (*RSS*), car un nœud mobile peut avoir un bon signal RSS associé à une faible bande passante ou à un coût élevé (Balasubramaniam et al., 2004). Il est clair que dans ce genre de circonstances, on s'attend à ce que la phase d'initiation puisse considérée davantage de paramètres de contexte tels que la bande passante, le trafic, le coût, la préférence de l'usager, etc. Cependant, l'initiation des relèves demeure préoccupante à cause de la multitude de paramètres de contexte à considérer d'une part et de l'hétérogénéité de ceux-ci d'autre part. Dans le cadre de cette thèse, on a utilisé la logique floue (*Fuzzy logique*) pour adresser certaines problématiques liées à la préparation des relèves. En effet, la logique floue est un puissant outil qui permet de créer de la précision à partir de l'imprécision (Zadeh, 1972).

La décision ou la sélection de la destination appropriée pour accomplir une relève dépend aussi bien du mobile que du réseau. Ainsi, on distingue trois stratégies de relèves: une relève contrôlée par le réseau (*network-controlled handoff*), une relève contrôlée par le mobile (*mobile-controlled handoff*) et une relève assistée (*mobile-assisted handoff*). Quant à l'exécution d'une relève, il peut s'accomplir par le biais des mécanismes de mobilité relevant de la couche réseau, transport ou application.

La qualité de service (QdS) est une notion qui peut prendre plusieurs significations dépendamment du contexte auquel on se réfère. D'une façon générale, elle est constituée de plusieurs métriques qui mesurent le degré de satisfaction d'un abonné ou d'un service. À titre d'exemple, les paramètres de QdS pour un mécanisme de mobilité peuvent être: la latence ou le délai de relève, le taux de perte des paquets, la signalisation, la probabilité de blocage, etc. Quant au choix d'un réseau de destination, les critères de QdS peuvent inclure: la bande passante, le temps de résidence, la variation du signal reçu (VRSS), la charge du réseau, etc.

Le temps d'interruption de service pendant lequel un nœud mobile ne peut ni envoyer ni recevoir des données durant une relève est désigné par délai de relève ou latence (*handoff delay / latency*).

Le taux de perte des paquets réfère aux paquets non reçus par le mobile pendant la période de relève.

La signalisation désigne l'ensemble des procédures de mise à jour des associations qui découlent de l'exécution d'une relève. Ces paramètres et bien d'autres tels que le taux de paquets réellement reçus (*throughput*) constituent les métriques de bases qu'essaient d'améliorer la plupart des travaux ayant trait à la gestion de mobilité. Dans le cadre des réseaux 4G, les critères de QdS auront une pondération encore plus importante dans les solutions de mobilité proposées vu que les utilisateurs solliciteront davantage les applications et les services multimédia.

1.2 Éléments de la problématique

Depuis l'apparition des systèmes de télécommunications sans fil, la demande des usagers en termes de mobilité et de QdS est en progression sans équivoque. Dans le but de satisfaire une telle demande, plusieurs solutions personnalisées ont été introduites dans le marché. Certaines de ces solutions ont comme objectif d'améliorer le débit et la connectivité tandis que d'autres misent sur la couverture et la portabilité des services. D'autres alternatives prônent la facilité du déploiement et un faible coût. Parmi tous ces efforts, aucune solution ne semble satisfaire les exigences des usagers. Ainsi, la future génération de réseaux mobiles s'annonce comme une sérieuse alternative qui devra, à priori, répondre massivement aux besoins des utilisateurs des technologies sans fil.

La conception et le déploiement de la 4G sous forme d'un nouveau système mobile sont loin d'être réalistes en raison des délais et des coûts importants qui seraient en jeu. En conséquence, la prochaine génération de réseaux mobiles consistera en l'intégration, de manière intelligente, des différents systèmes mobiles existants et ceux à venir. On s'attend donc à un environnement mobile constitué de différentes technologies sans fil exhibant l'interopérabilité et la complémentarité des services. Néanmoins, pour qu'un tel concept de réseau puisse se rendre au stade de l'exploitation réelle, de nombreux défis et problèmes doivent être résolus. Parmi ceux-ci, on retient, la gestion de mobilité, l'architecture d'intégration et la garantie de qualité de service.

Dans un environnement où coexistent une multitude de systèmes hétérogènes, les usagers mobiles auront souvent à exécuter des relèves aussi bien horizontales (*intra-technologie*) que verticales (*inter-technologie*). Par conséquent, il est primordial d'assurer la continuité des services indépendamment du mouvement et du réseau d'attache. En d'autres termes, il faut que les exigences des usagers en ce qui à trait à la qualité de service soient respectées lors du choix du réseau de destination (*handoff decision*) et de l'exécution des relèves (*handoff execution*). Les solutions de mobilité proposées dans la littérature peuvent être classées en trois catégories: mobilité au niveau réseau, mobilité au niveau transport et mobilité au niveau applicatif.

Dans le but de gérer la mobilité au niveau réseau, l'Internet Engineering Task Force (IETF) a proposé le protocole Mobile IPv6 (MIPv6) (Johnson et al., 2004) pour que les terminaux mobiles puissent maintenir une connexion active tout en se déplaçant à travers des systèmes homogènes ou hétérogènes. Cependant, MIPv6 est limité par le fait que le mobile doit mettre à jour ces informations à chaque fois qu'il change de position ou obtient une nouvelle adresse temporaire. Ceci induit d'importants délais pour les relèves ainsi qu'un gaspillage des ressources. D'autres améliorations ont été proposées dans le but de tenir compte de la mobilité locale et d'anticiper les relèves. Toutefois, aucune de ces nouvelles solutions ne permet d'avoir un faible trafic de signalisation, un délai de relève minimal et une perte de paquets tolérable (Pérez-Costa et al., 2003).

La mobilité au niveau transport vise à assurer une mobilité sans coupure (*seamless handoff*) tout en se désengageant des détails liés aux couches inférieures. Cependant, les solutions basées sur le protocole TCP (*Transmission Control Protocol*) nécessitent d'importants changements au niveau de l'architecture globale du protocole (Snoeren et al., 2000). En outre, TCP ne supporte pas l'utilisation simultanée de plusieurs interfaces sans fil (*multihoming*).

Avec la venue du protocole SCTP (*Stream Control Transmission Protocol*)(Stewart 2007) et plus précisément avec sa version mobile mSCTP (*mobile SCTP*) (Stewart et al., 2007), la mobilité au niveau transport a été relancée de façon intensive. Cependant, la mobilité basée sur mSCTP ne prend pas en considération la mobilité locale, ce qui induit des délais de relève additionnels. Autrement dit, les relèves basées sur mSCTP sont traitées de la même façon que si le mobile se déplace à l'intérieur d'un même domaine administratif. En outre, dans la spécification actuelle de SCTP, tous les accusés de réceptions sélectifs (SACKs) doivent être acheminés à la même source qui les a envoyés. En conséquence, durant la période de relève, tous ces accusés de réception seront perdus et le mécanisme de contrôle considérera les données associées à ces SACKs comme perdus et procédera à des retransmissions inutiles. Il est clair que ce genre de retransmissions, non sollicités, réduira de façon considérable le taux des paquets réellement reçus (*throughput*) pendant les périodes de relève.

Au niveau application, la mobilité est basée sur le protocole SIP (*Session Initiation Protocol*). Toutefois, la mobilité avec SIP présente encore d'importants délais de relève et de trafic de signalisation comparativement à la mobilité au niveau réseau surtout lorsqu'il s'agit d'une relève à travers un réseau UMTS (Banargee et al., 2004).

Par ailleurs, dans le cadre de la 4G, les mobiles se déplaceront de manière libre entre différentes plateformes mobiles. Ils auront donc à décider de leurs prochaines destinations à chaque fois qu'ils engagent une relève verticale. Traditionnellement, le choix de la destination se base uniquement sur la qualité du signal reçu (RSS). Dans un environnement multi-accès et multi-technologies, cette stratégie n'est plus appropriée car un réseau peut offrir une bonne qualité de signal RSS mais celle-ci peut être associée à une faible bande passante ou un coût élevé. Il est donc essentiel de disposer d'une stratégie de relève adaptée aux circonstances d'un milieu hétérogène, aux exigences des usagers et de la diversité des services offerts.

De plus, dans un contexte où les usagers mobiles engageront de plus en plus d'applications sensibles au délai, l'initiation des relèves demeure un enjeu prioritaire. En effet, afin d'éviter toute éventuelle dégradation de la QdS, le mobile doit être en mesure de décider quand et sous quelles conditions une relève devra être initiée. Toutefois, la définition de telles conditions nécessite un accès permanent aux informations de contexte relatives au réseau d'attache, et ce, pour détecter toutes anomalies relatives à la perturbation des services engagés. De plus, on devra tenir compte du fait que les informations de contexte peuvent être exprimées sous plusieurs formes (numérique ou linguistique). On s'attend également à ce que la phase de préparation soit en mesure d'identifier le type de relève qu'on doit considérer (forcée ou volontaire) afin de faciliter le choix du prochain point/réseau d'attache.

La présence accrue des réseaux et d'opérateurs mobiles rendra difficile l'obtention des informations de contexte dans la mesure où ceux-ci expriment une forte réticence au partage de leurs bases de données. De ce fait, la préparation des relèves devra inclure une analyse de contexte efficace qui tiendra compte de la diversité des technologies d'accès et de la confidentialité des réseaux du voisinage.

Afin de tirer profit de la complémentarité des services et de l'itinérance globale, une architecture d'intégration s'avère nécessaire. Actuellement, les efforts d'intégration déployés se basent sur deux modèles d'intégration proposés dans le cadre de l'intégration WLAN/3G. Les modèles en question concernent le couplage fort (*tight coupling*) et le couplage léger ou faible (*loose coupling*). Les architectures d'intégration susmentionnées présentent de sérieuses faiblesses. En ce qui concerne le couplage fort, le réseau WLAN est vu comme une extension du réseau 3G. Par conséquent, le trafic provenant du WLAN doit transiter par le réseau cœur 3G, ce qui entrainera une saturation de ce dernier. De plus, WLAN et 3G doivent appartenir au même opérateur. En ce qui concerne le couplage faible, les réseaux WLAN et 3G demeurent indépendants mais la qualité de service persistera en complète dépendance des conditions d'Internet. En effet, en l'absence de coordination directe entre le réseau d'attache et celui de destination, les relèves risquent d'être plus longues et induisent, en conséquence, une importante perte de paquets. En outre, les deux approches d'intégration citées plus haut, se restreignent seulement aux réseaux WLAN et 3G, ce qui laisse les réseaux émergents tels que les MANETs et les réseaux de capteurs loin des enjeux d'intégration envisagés par la 4G. Par ailleurs, une architecture d'intégration devra prendre en considération certaines exigences liées à l'évolutivité, au coût, à la facilité du déploiement et à la continuité des services offerts.

Il a été rapporté dans la littérature (Akyildiz et al., 2005) que la façon la plus directe et la plus simple pour assurer une liberté de mouvement, à travers différents systèmes mobiles, consiste à avoir des accords de services et d'itinérance (*service level agreement ou SLA*) entre chaque paire de réseaux. Cependant, cette option est loin d'être pratique lorsque le nombre de réseaux est important. De plus, les opérateurs de téléphonie mobile et les réseaux privés sans fil sont très réticents à l'idée d'autoriser l'accès à leurs données internes même si ce genre d'accès est parfois nécessaire pour compléter des opérations d'authentification ou de facturation. Une architecture d'intégration devra donc assurer l'itinérance globale des usagers mobiles sans avoir à utiliser des accords bilatéraux directs entre les réseaux intégrés.

1.3 Objectifs de recherche

L'objectif principal de cette thèse est de proposer des mécanismes de gestion de mobilité et d'intégration adaptés aux exigences des réseaux 4G en termes de connectivité, de qualité de service et d'interopérabilité. Plus précisément, cette thèse vise les objectifs suivants:

- Analyser les architectures d'intégration et les mécanismes de gestion de mobilité proposés dans la littérature afin d'en identifier les limites;
- Proposer de nouveaux mécanismes pour la gestion de mobilité de bout-en-bout, incluant l'initiation des relèves et la sélection des réseaux de destination;
- Concevoir une architecture d'intégration qui assure l'itinérance globale tout en respectant les exigences de la 4G en ce qui a trait à la mobilité et la continuité des services;
- Évaluer la performance des solutions proposées moyennant des modèles analytiques et des simulations en se comparant aux travaux antérieurs qui abordent des problématiques similaires.

1.4 Esquisse méthodologique

Dans le but d'atteindre nos objectifs de recherche de manière structurée, nous commencerons par une revue de littérature pertinente couvrant les travaux liés aux réseaux mobiles 4G. Tout au long de cette phase notre intérêt portera sur les travaux antérieurs ayant abordé des problématiques liées aux architectures d'intégration, à la gestion de mobilité et aux stratégies d'initiation et de décision des relèves. Cette façon de procéder nous permettra de déceler les exigences de la 4G en matière de connectivité, d'itinérance, de continuité des services et de support d'hétérogénéité.

Une fois qu'on a identifié les enjeux majeurs de la prochaine génération, nous allons commencer par aborder la problématique de mobilité en proposant un nouveau mécanisme de gestion de mobilité opérant au niveau transport et qui vise à améliorer la qualité de service et la transparence des relèves horizontales et verticales. En effet, l'approche de mobilité proposée est basée sur le nouveau protocole de transport communément appelé SCTP (Stream Control Transmission Protocol) ainsi que sur son extension mobile désignée par: *mSCTP*.

Ensuite, nous concevrons une nouvelle stratégie de préparation de relève. Celle-ci concerne l'initiation de relève et la sélection du réseau de destination le plus approprié. La phase d'initiation de relève fera appel à la logique floue qui permet de prendre en considération aussi bien les données précises (i.e., numériques) ainsi que celles qui ne le sont pas (i.e., linguistiques). De plus, le mécanisme d'initiation proposé assurera l'identification des relèves forcées de celles qui sont volontaires. Par ailleurs, l'approche de sélection introduite permettra à un usager mobile de choisir la destination qui répond le mieux possible à ses besoins de QdS ainsi qu'à ses préférences. Afin de garantir la confidentialité des informations de contexte relatives à chacun des réseaux intégrés, on se basera sur une analyse de contexte efficace. Celle-ci tiendra compte de l'hétérogénéité des réseaux ainsi que de la confidentialité des informations échangées.

En se basant sur les limites des architectures déjà introduites et sur les recommandations et spécifications d'organismes tels que l'IETF et le 3GPP/3GPP2, nous proposerons une nouvelle architecture d'intégration hybride basée IP. En d'autres mots, l'architecture proposée inclura les avantages des schémas d'intégration antérieurs, en l'occurrence le couplage fort/faible et l'utilisation d'une troisième autorité (*third party approach*). Tout au long de la conception de cette nouvelle architecture, nous ferons un compromis entre les différentes recommandations et exigences tels que : l'évolutivité, la sécurité, la facturation, l'itinérance, le coût, etc. Nous insisterons, en particulier, sur la réutilisation des entités existantes, la facilité du déploiement et la garantie d'une mobilité sans coupure.

L'analyse des performances des solutions proposées sera basée sur des simulations ainsi que sur des modèles analytiques. Les outils d'évaluation utilisés dans le cadre de cette thèse sont: ns-2 et MATLAB. Nous considérerons plusieurs scénarios de simulations et de tests numériques. Une multitude de métriques seront adoptées lors de cette analyse. Nous effectuerons également une étude comparative avec les autres propositions disponibles dans la littérature.

1.5 Principales contributions et originalité

Dans le cadre de cette thèse, trois principales contributions peuvent être considérées. La première contribution concerne la proposition d'un nouveau mécanisme de gestion de mobilité de bout-en-bout. La deuxième est liée à l'introduction d'une stratégie efficace de préparation des relèves. La troisième émane de la conception d'une nouvelle architecture hybride visant à intégrer différents systèmes mobiles autour d'une dorsale IP.

- Protocole hiérarchique de gestion de mobilité au niveau transport:

La mobilité au niveau transport basée sur le protocole mSCTP est considérée comme une alternative sérieuse pour les applications sensibles au délai (Iyengar et al., 2006). Toutefois, les solutions de mobilité proposées à ce niveau souffrent de certains problèmes tels que le délai de relève et les retransmissions non désirables (*spurious retransmissions*). En conséquence, nous proposons un nouveau protocole de gestion de mobilité basée sur SCTP/mSCTP et qui prend en considération la mobilité locale et globale. Cette façon de faire n'a jamais été expérimentée auparavant pour la mobilité au niveau transport. Les résultats montrent que notre approche permet de réduire les délais non désirables relatifs aux déplacements des mobiles à l'intérieur d'un même domaine administratif. De plus, les retransmissions inutiles de données dues à la perte des accusés de réception (SACK) lors des relèves sont contrôlées, de ce fait notre travail constitue une contribution originale.

- Stratégie de préparation de relève:

La préparation d'une relève est souvent négligée au détriment de son exécution. Et, même si elle est prise en considération, la qualité du signale reçu (RSS) demeure la principale métrique pour l'initiation et la décision des relèves. Dans le but de remédier à cette anomalie, nous introduisons une nouvelle stratégie de préparation de relève basée sur la logique floue. Elle permet, entre autres, de décider du moment opportun pour le déclenchement d'une relève ainsi que de la détermination de son type i.e., forcée ou volontaire. L'identification de la nature de la relève à initier est très importante pour la phase de sélection du réseau de destination. De plus, nous avons conçu une fonction de préférence adaptée aux environnements où coexistent plusieurs technologies sans fil. Cette fonction est utilisée pour sélectionner le réseau de destination le plus approprié en termes de stabilité et du respect de la préférence des usagers. En outre, la stratégie proposée incorpore une analyse de contexte efficace qui permet de garantir la disponibilité et la confidentialité des informations de contexte à travers des environnements hétérogènes.

- Architecture d'intégration pour les réseaux métropolitains

La stratégie de préparation des relèves et le protocole de mobilité proposés ont été mis à contribution moyennant une nouvelle architecture d'intégration évolutive et interopérable. L'architecture proposée permet la coexistence de plusieurs systèmes mobiles indépendamment des technologies qu'ils utilisent. Par ailleurs, elle assure l'itinérance globale des usagers mobiles en introduisant une tierce partie qui permet de réduire, de façon considérable, le nombre d'accords bilatéraux entre les réseaux intégrés. Par ailleurs, elle intègre d'importantes fonctionnalités dont le but est de garantir l'accès aux informations de contexte, l'authentification et la coopération avec les systèmes interconnectés. Il est à noter également que l'architecture proposée est couplée avec un mécanisme de mobilité qui permet d'assurer des relèves sans coupure tout en essayant de satisfaire au maximum les préférences des usagers.

1.6 Plan de la thèse

Le reste de la présente thèse sera répartie comme suit : le chapitre 2 présente une revue de littérature exhaustive et critique des principaux enjeux de la prochaine génération des réseaux mobiles à savoir: l'intégration et la gestion de mobilité. Ensuite, les chapitres 3 à 5 contiennent les différents articles relevant de cette thèse. Le chapitre 6 inclut une discussion générale des résultats obtenus. Enfin, le chapitre 7 présente une récapitulation des travaux ainsi que des recommandations sur les travaux futurs.

Plus précisément, le chapitre 3 intitulé : *Adaptive end-to-end mobility scheme for seamless horizontal and vertical fhandoffs* est un article accepté pour publication dans *Ubiquitous Computing and Communication Journal.* Dans ce chapitre, nous présentons un nouveau protocole de gestion de mobilité de bout en bout qui vise à réduire les délais de relève et les pertes de paquets lorsque le nœud mobile se déplace au sein d'un même domaine administratif. De plus, cette solution permet de palier le problème de détérioration du flux de données reçu durant la période de changement du point d'attache.

Le chapitre 4 intitulé : *A Decision Making Strategy for Horizontal and Vertical Handoffs* est un article soumis à la revue *Journal of Computers (JCP Academy Publishers*. Dans cet article, nous proposons une stratégie de décision de relève qui a comme finalité d'assurer une préparation appropriée des relèves. Plus spécifiquement, nous proposons une solution efficace à la problématique d'analyse de contexte dans un milieu hétérogène. De plus, un mécanisme d'initiation de relève basé sur la logique floue et un processus de sélection des réseaux de destination ont été également proposés.

Le chapitre 5 désigné par *Towards Cross Layer Mobility Support in Metropolitan Networks* est un article accepté pour publication à la revue *Computer Communications* (*Elsevier*). Dans ce chapitre, nous proposons une architecture d'intégration qui se veut adaptée aux exigences de la prochaine génération de réseaux mobiles.

Le chapitre 6 est une discussion générale relative aux résultats obtenus ainsi qu'une synthèse des travaux réalisés. Finalement, le chapitre 7 dresse un bilan des travaux accomplis par rapport à nos objectifs de recherche. De plus, nous y rapportons les limites de nos contributions ainsi que les éventuels extensions et recommandations pour les travaux futurs.

CHAPITRE 2 PROBLÉMATIQUES D'INTEGRATION DES RÉSEAUX 4G

De nos jours, les systèmes de communication sans fil sont omniprésents et offrent des services complémentaires et diversifiés. Afin de tirer profit de cette diversité et de créer une nouvelle valeur ajoutée pour ce genre de réseaux mobiles, l'intégration et l'interopérabilité de ces systèmes semblent être le choix le plus approprié. Cependant, une multitude de problèmes et de défis émanent de cette intégration. Parmi ceux-ci, la conception de nouvelles architectures d'intégration, la gestion de mobilité, la sécurité, la facturation et la garantie d'une meilleure connectivité constituent les enjeux majeurs de la 4G. Dans le but de faire face aux défis susmentionnés, plusieurs travaux ont été énumérés dans la littérature. Toutefois, les solutions proposées présentent encore des problèmes et des faiblesses. Dans ce chapitre, nous allons effectuer une analyse approfondie et exhaustive des différents travaux relatifs aux défis préalablement cités. Plus spécifiquement, notre intérêt portera sur les architectures d'intégration existantes, la gestion de mobilité et les mécanismes de décision des relèves.

2.1 Architectures et approches d'intégration

À peine rentrée en marché, la troisième génération de réseaux mobiles (3G) se voit déjà dans l'impossibilité de répondre aux besoins grandissants de ses abonnés en ce qui concerne le débit, la connectivité, la sécurité et la qualité de service. En revanche, de nouvelles solutions sans fil émergent et offrent des débits plus importants et une couverture encore plus large. Devant cette réalité, le besoin d'une nouvelle génération de réseaux mobiles devient de plus en plus plausible.

Dans la pratique, deux approches sont possibles pour réaliser cette nouvelle génération de réseaux mobiles (4G). La première consiste à concevoir, à l'instar des générations précédentes, un nouveau système sans fil incluant une partie cœur (réseau filaire) et un réseau d'accès. La deuxième approche vise à intégrer, de manière intelligente, les réseaux mobiles existants et ceux à venir dans le but de permettre aux usagers mobiles d'être toujours servis par le meilleur réseau disponible (Akyildiz et al., 2005). La première option est visiblement loin d'être réaliste puisqu'elle nécessite une nouvelle technologie et un nouveau déploiement, ce qui risque d'être coûteux en termes de coût et de délai de réalisation. La seconde option est beaucoup plus réaliste et semble susciter davantage l'intérêt de la communauté scientifique.

Toutefois, plusieurs problèmes et défis résultent de cette intégration. À titre d'exemple, on peut citer la gestion de mobilité à travers des environnements hétérogènes, la garantie de qualité de service et la conception d'une architecture d'intégration appropriée. D'une manière générale, on s'attend à ce que la future génération de réseaux mobiles soit:

- économique: c'est-à-dire il faut que les coûts de déploiement et de mise en service soient abordables;
- évolutive: en d'autres termes, elle doit être ouverte et flexible en vue de l'ajout de nouveaux composants;
- assure une mobilité sans coupure: autrement dit, elle doit garantir des relèves transparentes et avec un minimum de délai d'interruption et de perte de paquets;
- sécuritaire, c'est-à-dire elle doit faire prévaloir la sécurité lors de l'itinérance des usagers mobiles à travers des systèmes sans fil hétérogènes;
- garantit une connexion permanente au meilleur réseau disponible.

Dans ce sens, de nombreux efforts d'intégration ont été initiés afin d'assurer l'interopérabilité entre les différentes technologies. La grande majorité des solutions proposées se contentent de l'intégration des réseaux WLAN et 3G et se justifient par la complémentarité des services qu'ils offrent les uns par rapport aux autres. Pour ce faire, deux principales architectures ont été proposées pour intégrer les réseaux WLAN et 3GPP/3GPP2. Les architectures en question sont communément appelées: couplage fort et couplage faible.

2.1.1 Couplage fort

La Figure 2.1 décrit de façon sommaire l'architecture du couplage fort (*tight coupling*). Cette architecture a été introduite pour intégrer les réseaux WLAN et 3GPP/3GPP2. Dans ce genre de scénario, le réseau WLAN apparaît comme une extension du réseau 3G. Plus précisément, le WLAN est directement connecté au réseau cœur 3G. Il est donc nécessaire de garantir une transparence mutuelle entre les réseaux WLAN et 3G. Ceci peut être réalisé moyennant des protocoles déployés au niveau du point d'ancrage (point d'interconnexion) des réseaux. Ainsi, la mobilité des abonnés entre un WLAN et un 3G sera basée sur les protocoles de gestion de mobilité de 3GPP/3GPP2. De plus, le trafic provenant du WLAN transitera à travers le réseau 3G avant d'être acheminé vers l'extérieur. Cette façon de faire causera de sérieux problèmes au réseau 3G puisque celui-ci n'est pas conçu pour supporter un trafic haut débit. Par ailleurs, les réseaux WLAN et 3G doivent appartenir au même opérateur pour faciliter le déploiement et garantir la sécurité. En outre, les interfaces sans fil WLAN des nœuds mobiles devraient implémenter la pile des protocoles du système 3G. Ceci risque d'être coûteux en termes de portabilité et de flexibilité.

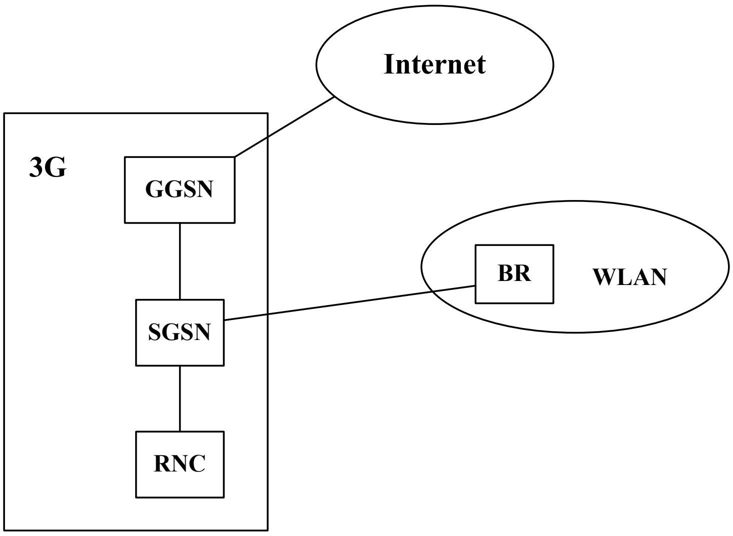


Figure 2.1 Architecture d'un couplage fort

2.1.2 Couplage faible ou léger

L'architecture d'un schéma de couplage faible (*loose coupling*) est illustrée à la Figure 2.2.

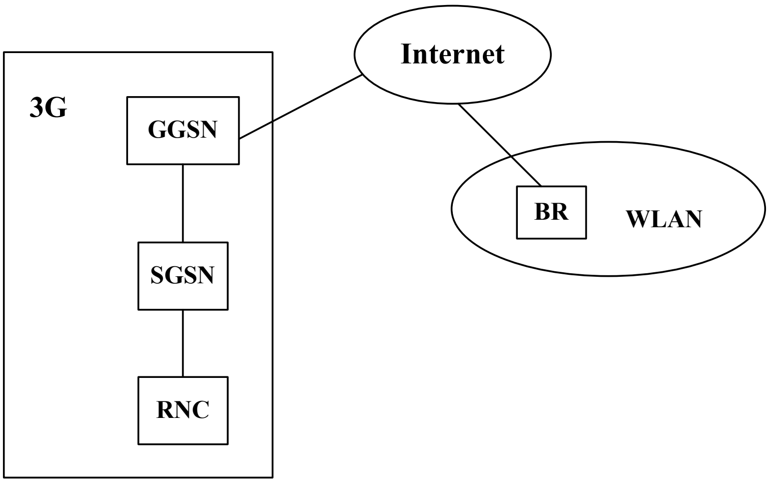


Figure 2.2 Architecture d'un couplage faible

Dans ce cas, les réseaux WLAN et 3G sont indépendamment connectés au réseau externe de données (réseau IP). Autrement dit, le trafic provenant du réseau WLAN ne sera plus contraint à passer par le réseau 3G pour accéder à l'extérieur. En conséquence, les réseaux WLAN et 3G peuvent appartenir à des opérateurs indépendants et n'auront pas à se soucier de la compatibilité de leurs technologies. De plus, la gestion de mobilité, l'authentification et la facturation se basent en grande partie sur les protocoles proposés dans le cadre de l'IETF. Cette approche semble être appropriée pour intégrer des systèmes mobiles homogènes (même technologie) et hétérogène (technologies différentes). Cependant, pour assurer un libre mouvement des usagers mobiles, des accords multilatéraux entre les réseaux intégrés sont exigés. Par ailleurs, afin d'assurer une mobilité sans coupure, les protocoles de mobilité relevant de chaque réseau doivent être interopérables. De plus, il est souvent difficile de garantir la qualité de service dans ce genre de scénario, car celle-ci dépend, en grande partie, des conditions des réseaux externes. Par exemple, la latence et la perte de paquets lors d'une relève verticale sont généralement plus importantes dans ce genre d'architecture.

2.1.3 Couplage hybride

Les scénarios d'intégration faible (*loose coupling*) et fort (*tight coupling*) ne font pas l'unanimité et chacune de ces solutions présente des avantages et des inconvénients. Devant ce constat, une solution hybride semble être un choix judicieux. Dans ce sens, Wang et al. (2001) proposent d'intégrer chaque paire de réseaux adjacents moyennant une passerelle. Cependant, ce type d'intégration n'est pas pratique car il nécessite un composant entre chaque paire de réseaux. De plus, les auteurs supposent l'existence d'accords bilatéraux entre les réseaux intégrés, ce qui risque de devenir fastidieux surtout lorsque le nombre des systèmes à intégrer est élevé. Havigna, et al. (2001) ont introduit une nouvelle architecture qui vise à traiter séparément le trafic de signalisation et celui des données. Toutefois, cette solution nécessite un nombre considérable de nouveaux composants réseaux pour gérer la signalisation et le flux de données. Il est clair que ce genre de proposition est très coûteux du point de vue infrastructure et déploiement. Salkintziz et al. (2002) ont présenté deux scénarios d'intégration basés sur le couplage faible et fort pour intégrer les réseaux WLAN et UMTS. Ces scénarios utilisent deux nouvelles entités désignées par GPRS Interworking Function (GIF) et WLAN Adaptation Function (WAF) afin d'assurer la correspondance des fonctionnalités entre les deux réseaux. Buddhikot et al. (2003) ont proposé une architecture fondée sur le couplage faible pour intégrer les réseaux cdma2000 et IEEE 802.11. Cependant cette catégorie d'intégration est spécifique aux réseaux susmentionnés et hérite également des inconvénients du couplage faible. Dans la proposition de (Song et al., 2003), le trafic sensible au délai est acheminé en utilisant le couplage fort, tandis que le trafic normal est acheminé par le biais d'un schéma d'intégration basée sur le couplage faible. Toutefois, plusieurs difficultés et limitations émanent de cette différenciation du trafic. Akyildiz et al. (2005) ont introduit une tierce partie dans sa proposition d'intégration qui demeure en complète dépendance du couplage faible. Le but est d'éviter les accords multilatéraux et d'adresser certaines problématiques liées à l'authentification et à la facturation dans les environnements hétérogènes. Pour ce faire, on a introduit deux unités désignées par : NIA (Network Interoperating Agent) et IG (Interworking Gateway). Toutefois, cette approche ne permet pas de garantir la qualité de service car l'approche de mobilité proposée avec cette architecture ne considère que le signal RSS pour choisir les prochains réseaux de destination. De plus, le délai de relève dépend de la localisation du NIA. Dans (Makaya et al., 2007), l'intégration est également basée sur une tierce partie et sur le couplage faible moyennant l'entité IDE (Interworking Decision Engine). Cependant, cette solution exclut intégralement les avantages du couplage fort puisqu'elle demeure entièrement basée sur le couplage faible. En plus, la garantie de qualité de service pour cette proposition dépend des conditions de la dorsale IP et de la position du nœud correspondant.

2.2 Stratégies de décision de relèves

Dans les réseaux mobiles de prochaine génération, les usagers seront souvent amenés à exécuter des relèves verticales. Ce scénario est d'autant plus réaliste que les terminaux mobiles seront équipés d'interfaces radio multi-accès (McNair et al., 2004). Parmi les problématiques qui demeurent encore ouvertes dans ce nouveau contexte de mobilité, on peut considérer celle qui vise à garantir à un nouvel abonné d'être toujours connecté au meilleur réseau disponible (*Always Best Connected ABC*) (Gustafsson et al., 2003). En effet, la majorité des mécanismes de mobilité disponibles dans la littérature ne peuvent pas, en tout temps, garantir le concept *ABC* puisqu'ils se basent généralement sur la qualité du signal reçu (RSS) pour choisir le prochain réseau d'attache. Ce genre de critère peut être suffisant pour des relèves homogènes. Par contre, dans le cas de relèves verticales, il est clair qu'un réseau de destination peut très bien avoir un bon signal RSS mais celui-ci peut être également associé à une faible bande passante ou à un coût monétaire élevé. De plus, pour optimiser le processus de relève, on doit choisir le moment et les conditions appropriées pour initier une relève. En conséquence, pour qu'un mécanisme de mobilité soit en mesure de gérer les relèves verticales dans un environnement 4G, il doit incorporer des procédures de décision de relèves tenant compte aussi bien de l'initiation des relèves que de la sélection des réseaux de destination. Dans cette section, nous allons présenter une revue de littérature des stratégies de relèves les plus connues afin d'en déceler les avantages et les faiblesses. Plus précisément, les stratégies de relèves existantes peuvent être classées en quatre principales catégories: fonction de décision, stratégies à attributs multiples, logique floue et analyse de contexte.

2.2.1 Stratégies basées sur une fonction de décision

Dans ce genre de solutions, la fonction de décision permet de mesurer la satisfaction de l'usager par rapport au choix du réseau de destination. Une telle fonction est exprimée sous forme d'une somme de facteurs de coûts associés à chacun des services requis. De plus, la préférence envers chaque service est quantifiée par un poids qui désigne une valeur comprise entre 0 et 1.

La première fonction de coût utilisée dans le cadre d'une stratégie de relève a été introduite dans (Wang et al., 1999). Les paramètres considérés dans cette fonction sont: la bande passante, la consommation d'énergie et le coût monétaire. Toutefois, cette fonction ne peut pas répondre aux exigences d'un environnement 4G vu qu'elle considère un nombre limité de paramètres de contexte. Une autre fonction d'utilité a été proposée par Chen et al. (2004). Celle-ci se base essentiellement sur deux paramètres, à savoir la bande passante et la vitesse du mobile. Dans ce sens, Guo et al. (2005) ont proposé une stratégie de décision de relève qui vise à évaluer le réseau de destination moyennant une fonction de coût. Celle-ci se propose de trouver un compromis entre la satisfaction de l'usager et la qualité du réseau de destination. Cependant, dans toutes les solutions susmentionnées, on se contente de supposer la disponibilité des informations de contexte. De plus ce type de fonction ne permet pas de prévoir la stabilité du réseau choisi. En d'autres termes on ne peut pas savoir si le réseau de destination qu'on a choisi, ne va pas perdre la qualité de ces paramètres de contexte après une courte durée. C'est le cas, par exemple, avec les réseaux très dynamiques tels que les réseaux ad hoc ou MANETs.

2.2.2 Stratégies de décision à attributs multiples (*Multiple Attribute Decision Strategies*)

La décision de relève est un cas particulier des problèmes connus dans la littérature sous le nom de: choix multicritères ou décision multi-objectifs. En effet, le problème de décision de relève consiste, entre autres, à choisir une meilleure destination parmi une liste de réseaux candidats. Parmi les méthodes les plus populaires pour résoudre ce genre de problèmes on peut citer:

- la méthode SWA (*Simple Additive Weighting*) qui est considérée comme la plus simple et la plus utilisée pour effectuer des choix multicritères. Plus précisément, elle consiste à définir une fonction de coût comme suit: $f_n = \sum_i \omega_i^{n,s} \cdot C_i^{n,s}$

où

$C_i^{n,s}$: réfère au coût/préférence du service *s* sous le critère *i* dans le réseau *n*,

$\omega_i^{n,s}$: réfère au poids de coût/préférence du service *s* sous le critère *i* dans le réseau *n*,

La meilleure solution consistera à trouver le réseau qui satisfait : $\underset{n}{Min}(f_n)$ ou $\underset{n}{Max}(f_n)$ selon qu'on cherche à minimiser le coût ou à maximiser la préférence.

- la méthode TOPSIS (*Technique for Ordering Preference by Similarity Ideal Solution*) permet de choisir le réseau de destination le plus proche de la solution idéale et le plus loin de la pire solution (Young-Jou, et al., 1994).
- la méthode AHP (*Analytic Hierarchy Process*) consiste à décomposer le problème de sélection du réseau de destination en plusieurs sous-problèmes et affecte des poids à chacun d'entre eux. Pratiquement, AHP peut être décomposé en un processus de trois étapes (Saaty et al., 1990) :

a) décomposer le problème de décision en plusieurs niveaux de hiérarchie (identification des critères de décision);

b) comparer chaque facteur aux autres à l'intérieur du même niveau;

c) calculer la somme du produit des poids obtenus à partir des différents niveaux. La solution correspond à celle ayant la plus grande somme.

- la méthode GRA (*Grey Relational Analysis*) est utilisée pour classifier les réseaux candidats et pour choisir celui ayant le plus grand degré ou rang (*rank*).

Dans (Quiqyang et al., 2005), on a proposé un mécanisme de sélection de réseau combinant les méthodes AHP et GRA dans le but d'arriver à un compromis entre la préférence de l'usager, les services d'applications et les conditions du réseau. Toutefois, cette approche demeure complexe et ne tient pas compte des données imprécises qui constituent une bonne partie des critères de contexte. De plus, l'analyse et la gestion du contexte ne sont pas considérées dans ce genre de solutions.

2.2.3 Stratégies basées sur la logique floue et les réseaux de neurones

Afin de remédier aux problèmes d'utilisation des données imprécises dans le cadre des stratégies à attributs multiples, on a introduit le concept de logique floue et des réseaux de neurones dans les stratégies de sélection. Dans ce sens, Pahlavan et al. (2000) ont développé un algorithme de relève verticale basée sur les réseaux de neurones pour satisfaire la bande passante des usagers. Toutefois, ce genre d'algorithme est loin d'être approprié pour les composants mobiles ayant des capacités limitées. De plus, cette solution nécessite une préparation préalable des réseaux de neurones, ce qui est coûteux en termes de temps et de consommation des ressources. Dans (Chi-Hsing et al., 1999) les auteurs ont proposé une solution utilisant la logique floue et qui permet à un usager de faire son choix entre un réseau terrestre et un réseau satellitaire. Toutefois, cette solution reste spécifique aux réseaux susmentionnés. Makela et al. (2000) ont utilisé les concepts de logique floue et des réseaux de neurones pour prendre des décisions de relèves. Cependant, cette approche ne prévoit pas de mécanisme d'analyse de contexte pour garantir l'accessibilité et la disponibilité des informations de contexte.

2.2.4 Stratégies basées sur l'analyse du contexte

Ce genre de stratégie de décision de relèves se base sur les informations de contexte relatives au terminal mobile ainsi que sur celles relatives aux réseaux candidats. Ceci permet d'engager des décisions de relèves adaptées aux exigences et aux préférences de l'environnement hétérogène du nœud mobile (Wei et al., 2006).

Dans (Ahmed et al., 2006) on a développé et analysé un algorithme de décision de relève pour chaque type de service s'exécutant sur le mobile. Balasubramaniam et al. (2004) ont introduit un cadre décisionnel (*Framework*) incluant une catégorisation du contexte et un algorithme de décision de relève basé sur la méthode AHP. Toutefois, cette solution est basée sur une collecte centralisée des informations de contexte. Plus précisément, les informations du contexte sont gérées par un point unique (*Repository*) ce qui risque d'être fatal en cas de panne. De plus, celui-ci nécessite des communications fréquentes entre le terminal mobile et le réseau, ce qui entraîne une augmentation considérable de la charge sur le lien sans fil (*wireless link*).

2.3 Mécanismes de gestion de mobilité

La problématique de gestion de mobilité demeure un enjeu majeur et décisif pour la prochaine génération de réseaux sans fil. Dans le but de faire face à ce grand défi tout en respectant les exigences et les recommandations de la 4G, plusieurs solutions ont été proposées. Dans la pratique, celles-ci se réfèrent aux différentes couches de la pile de protocole TCP/IP.

2.3.1 Mobilité au niveau application

La mobilité au niveau application a suscité également beaucoup d'attention vu que ce genre d'approches est pratiquement indépendant des couches inférieures. Cette catégorie de solutions est basée sur le protocole SIP (Session Initiation Protocol). Ainsi, lorsqu'un nœud mobile se déplace à travers différents réseaux durant une session, il obtient d'abord une nouvelle adresse IP de la destination à visiter. Ensuite, il envoie un nouveau "*Session Invitation*" à son nœud correspondant (CN). Toutefois, lors d'un mouvement, le protocole SIP ne peut pas garantir le maintien d'une session TCP ou assurer la mise en correspondance des ports UDP. Pour remédier à ce handicap, d'autres extensions telles que S-SIP (Zhang et al., 2007) ont été proposées. Néanmoins, le principal inconvénient de cette méthode demeure le délai et le trafic de signalisation.

2.3.2 Mobilité au niveau réseau

Au niveau de la couche réseau, Perkin et al. (2002) ont proposé le protocole de mobilité le plus populaire, à savoir Mobile IPv4. Cependant, ce protocole souffre de faiblesses telles que la latence et la signalisation. Afin de remédier à certaines de ses limites, plusieurs extensions ont succédé à MIPv4. Ces améliorations concernent, en particulier, l'introduction de Mobile IPv6 (Johnson et al., 2004) en vue de réduire le trafic de signalisation. HMIP (Soliman et al., 2005) est une autre amélioration de Mobile IP qui se propose de gérer aussi bien la mobilité locale que globale (i.e., micro et macro mobilité). FMIP (Koodli, 2005) est également une extension de Mobile IP qui vise à assurer l'anticipation des relèves en vue de réduire la latence. La combinaison de ces deux dernières approches a donné naissance à F-HMIP (Jung et al., 2005) qui a comme objectif d'anticiper les relèves tout en gérant la mobilité locale. De plus, une multitude d'autres solutions ne relevant pas de l'IETF sont également disponibles dans la littérature. Toutefois, ces propositions demeurent encore coûteuses en terme de signalisation et ne garantissent pas la QdS à travers des environnements hétérogènes.

2.3.3 Mobilité au niveau transport

La gestion de mobilité au niveau transport vise à limiter la dépendance aux couches inférieures et à tirer profit des facilités de connexion et de contrôle de flux offertes à ce niveau. Dans ce sens, une nouvelle architecture a été proposée dans (Snoeran et al., 2000) et (Hsieh et al., 2003) pour adapter TCP à supporter de la mobilité. Toutefois, d'importants changements, au niveau de l'architecture du réseau, sont nécessaires pour atteindre cet objectif. Une autre solution basée sur TCP a été introduite dans (MALTZ et al., 1998). Cette proposition ne requiert pas des changements à l'infrastructure de la couche réseau. Cependant, elle présente une latence et une perte de paquets élevées. Avec la venue du protocole SCTP (*Stream Transmission Control Protocol*) proposé par Stewart (2007) et sa version mobile introduite par Stewart et al. (2007), le défi d'une mobilité au niveau transport est relancé, cette fois, avec plus d'assurance sur les chances de sa concrétisation.

Vu que les mécanismes de mobilité proposés dans le cadre de cette thèse se basent, en partie, sur le protocole SCTP, nous donnerons dans ce qui suit une brève présentation de ce protocole.

2.3.4 Introduction du protocole SCTP

Les protocoles de transport les plus populaires demeurent TCP (*Transmission Control Protocol*) et UDP (*User Datagram Protocol*). Toutefois, afin de répondre aux nouvelles exigences des applications en termes de fiabilité, de connectivité et de sécurité, l'IETF (*Internet Engineering Task Force*) a standardisé un nouveau protocole de transport qui se veut mieux adapté aux besoins des applications émergentes. Le protocole en question est le SCTP. Il a été conçu au départ pour résoudre le problème de transport de signalisation des applications de voix sur IP. Une première version a été proposée en octobre 2000 par le groupe de travail IETF SIGRAN (*IETF Signaling Transport*). À la suite aux intéressantes applications de SCTP comme une sérieuse alternative à TCP, un nouveau groupe de travail a été créé en Février 2001 (*TSVWG : Transport Area Working Group*) pour ajouter d'autres fonctionnalités à ce nouveau protocole. Dans la plus récente version du standard SCTP introduit par Stewart (2007), on a gardé les principaux points forts de TCP tel que le contrôle de flux, la détection des erreurs et la retransmission, tandis que de nouvelles propriétés ont été ajoutées telles que le *multihoming*, le *multistreaming* et la fiabilité partielle.

2.3.4.1 Association

Dans le but d'échanger des informations entre deux hôtes d'un réseau, on doit tout d'abord établir une relation/connexion entre ceux-ci. Dans la terminologie SCTP, cette relation est appelée *association* et les éléments qui communiquent entre eux sont appelés points terminaux. À l'instar de TCP et UDP, le protocole SCTP se base sur la notion de numéros de ports pour identifier les applications de la couche supérieure. Dans le cadre de SCTP, deux types de points terminaux peuvent être considérés: *single-homed* et *multi-homed*. Dans ce sens, un point terminal *sigle-homed* est défini par: [adresse IP, port SCTP], tandis qu'un point terminal *multi-homed* est défini par: [Adresse IP1,…, Adresse IPn, port SCTP].

- Établissement d'une association

L'établissement d'une association (connexion SCTP) se fait en quatre étapes comme le montre la Figure 2.3. Toutefois, l'échange des données peut être déjà amorcé dans le message *COOKIE-ECHO*. Ceci permet d'accélérer l'échange des données utiles pendant la phase d'établissement d'une association. De plus, pour pallier le problème de demi-connexion ou SYN attaques, le serveur ne doit réserver aucun espace mémoire avant que l'association ne soit complètement établie. Plus précisément, le processus d'initiation d'une association commence par l'envoi d'un c*hunk* (message) *INIT* pour inviter le serveur à démarrer le processus d'établissement d'association. Ensuite, le serveur répond par un *INIT-ACK* qui contient un cookie pour permettre aux associations SCTP de se munir contre les attaques SYN. Par la suite, le client renvoie le cookie au serveur par l'intermédiaire du *chunk COOKIE-ECHO*. À partir de ce moment, le serveur peut réserver des ressources au client et confirme l'établissement de l'association par le *chunk COOKIE-ACK*. Lors de l'établissement d'une association SCTP, les points terminaux (i.e., client & serveur) définissent un seul chemin primaire et un ou plusieurs chemins secondaires. Le chemin primaire sera utilisé pour le transfert des données tandis que les chemins secondaires seront utilisés pour la signalisation, les retransmissions ou comme chemin de récupération en cas de panne du chemin primaire.

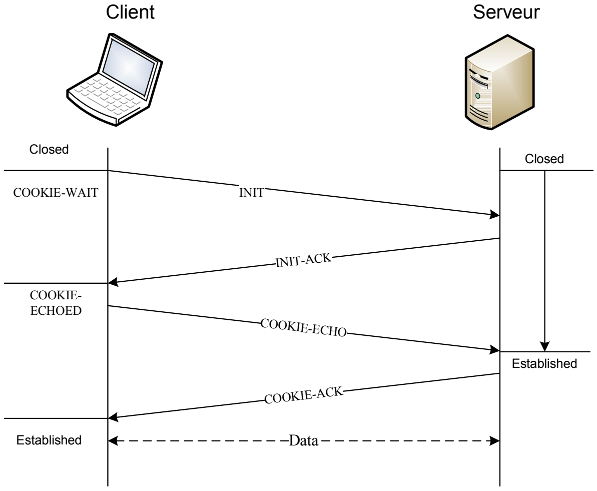


Figure 2.3 Établissement d'une association SCTP

- Fermeture d'une association

Afin de terminer une association, SCTP utilise soit le *graceful shutdown* ou *l'abortive shutdown*. Avec le *graceful shutdown*, SCTP permet une fermeture en trois étapes comme le montre la Figure 2.4. Avec *l'abortive shutdown*, la connexion est supposée être terminée de façon forcée ou suite à un événement inattendu.

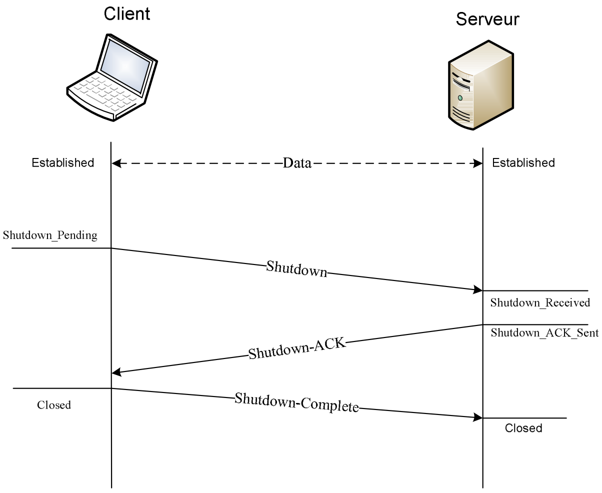


Figure 2.4 Fermeture en trois étapes d'une association SCTP

2.3.4.2 Multi-streaming

Le multi-streaming ou le multiplexage de flux consiste en la livraison de différents flux de données en séparant le transfert fiable des données du mécanisme de livraison. Ceci permet de s'adapter aux besoins spécifiques des applications utilisant SCTP. En effet, certaines applications peuvent avoir besoin seulement d'une remise en ordre partielle tandis que d'autres pourraient se contenter d'un transfert fiable qui ne garantit aucun ordre de transmission. De plus, il permet, selon Scharaf et al. (2006), à une association SCTP d'éviter le "head-of-line blocking" qui se produit lorsque plusieurs flux de données surviennent de façon indépendante dans une même association SCTP.

2.3.4.3 Multi-homing

La propriété du multi-homing est l'un des principaux atouts qui distinguent SCTP des autres protocoles. Cette propriété permet à un point terminal d'être atteint par plusieurs adresses IP. De plus, avec l'apparition de la version mobile de SCTP désignée par: mSCTP (Stewart et al., 2007), le multi-homing permet une grande flexibilité quant à l'essai de la mobilité au niveau transport.

Toutefois, dans la spécification actuelle de SCTP, le multi-homing est utilisé uniquement pour les retransmissions et non pas pour les envois simultanés des données.

2.3.4.4 Contrôle de flux et de congestion

Dans un système de communication, la congestion peut apparaître soit du côté récepteur soit dans le réseau. Du côté récepteur, la congestion est généralement due à la taille des filles d'attentes de réception tandis que dans le réseau, elle est couramment due à la saturation des liaisons. Dans le premier cas, le problème de congestion est résolu par le champ : *Advertised Receiver Window* (*a_rwnd*) qui se trouve dans les chunks de type *INIT*, *INIT-ACK* et *SACK*. Ce paramètre indique à l'émetteur combien de *bytes* le récepteur est prêt à recevoir. Dans le deuxième cas, le contrôle de flux se fait moyennant les mêmes algorithmes utilisés par TCP, en l'occurrence, le *Congestion Window* (*cwnd*) qui permet de contrôler le nombre de *bytes* que l'émetteur peut envoyer et le *Slow Start Threshold* (*ssthresh*) qui permet de choisir le bon algorithme de congestion au bon moment.

En guise de conclusion de cette sous-section, le tableau ci-dessous donne une récapitulation sommaire des propriétés relatives aux principaux protocoles de transport présentés dans la littérature.

Tableau 2.1 Comparaison des différents protocoles de transport

	UDP	TCP	SCTP
Démarrage	Aucun établissement de connexion	Établissement de la connexion en trois temps	Établissement de l'association en 4 temps, avec échange de cookies, l'échange de data peut être effectué dès le 3ème message d'établissement (dans COOKIE-ECHO et COOKIE-ACK)
Fiabilité	Aucun acquittement des messages, aucune assurance de livraison	Acquittement des messages pour assurer la transmission. L'acquittement sélectif est optionnel dans TCP	Acquittement des messages pour assurer la transmission. L'acquittement est sélectif; seuls les messages erronés sont retransmis
Ordre de livraison	Aucun mécanisme pour assurer l'ordre des messages	Numérotation des messages	Les messages peuvent être ordonnés ou non ordonnés. Même les messages non ordonnés gardent une assurance de livraison
Contrôle de la congestion	Aucun mécanisme	Mécanisme (Additive Increase, Multiplicative Decrease)	Mise en œuvre des mécanismes de contrôle de congestion (Additive Increase, Multiplicative Decrease)
Procédure de fermeture	Aucune	Fermeture de la connexion spécifiée. Le mode " half-closed" est possible	Fermeture de l'association spécifiée. Le mode " half-closed" est possible
SYN attack	Insensible (aucune connexion)	C'est un des problèmes de TCP	Résolu avec le Cookie
Head-of-Line Bloking	Insensible	Partiellement résolu en ouvrant plusieurs connexion	Résolu car SCTP permet plusieurs streams (non bloquants entre eux) dans la même association
Multi-homing	Pas supporté	Pas supporté	Supporté

2.3.5 Gestion de mobilité avec SCTP/mSCTP

Avec l'apparition de SCTP et en particulier avec sa version mobile: mSCTP, la mobilité au niveau transport est devenue plus que jamais une sérieuse alternative. Celle-ci se propose de garantir des relèves transparentes et sans coupures lorsqu'un mobile se déplace au sein d'un même réseau ou à travers des réseaux hétérogènes.

Dans ce sens, Ma et al. (2004) ont introduit une approche exploitant la propriété du multi-homing et l'adressage dynamique d'une association pour assurer des relèves verticales sans coupure entre les réseaux UMTS et WLAN. Fu et al. (2004) ont également proposé un cadre de mobilité pour assurer une itinérance sans coupure entre les réseaux UMTS, WLAN et satellitaire. Koh et al. (2004) ont proposé d'intégrer de nouvelles règles d'initiation d'associations SCTP afin d'améliorer le flux des données reçues (*throughput*) durant l'exécution des relèves. Dans le but de tirer un grand profit de la fonctionnalité du multi-homing en ce qui concerne l'échange des données entre deux points terminaux, Iyengar et al. (2006) et Fracchia et al. (2007) ont proposé d'incorporer au standard SCTP de nouvelles techniques de transmissions simultanées des données. Toutefois, les solutions proposées jusqu'à date demeurent coûteuses en termes de latence et de signalisation vu qu'elles ne tiennent pas compte de la mobilité locale. Par ailleurs, certains effets cachés tels que les acquittements perdus (SACK) pendant la phase de relève se traduisent par une réduction du flux des données reçues (*throughput*) durant les phases de relève.

CHAPITRE 3

ADAPTIVE END-TO-END MOBILITY SCHEME FOR SEAMLESS HORIZONTAL AND VERTICAL HANDOFFS

Abdellatif Ezzouhairi, Alejandro Quintero, Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
P.O. Box 6079, succ. Centre-Ville, Montreal, Quebec, H3C 3A7, Canada
Phone: (514) 340-3240 ext. 4685. Fax: (514) 340-3240
E-mail: {Abdellatif.Ezzouhairi; Alejandro.Quintero; Samuel.Pierre}@polymtl.ca

Abstract

Mobility management constitutes one of the most significant task to be investigated for Next Generation Mobile Networks (4G). Motivated by connectivity facilities and flow control offered at the transport layer, a number of Stream Control Transmission Protocols (SCTPs) based mobility schemes have been proposed to handle this important issue. However, these proposals are hindered by drawbacks such as unnecessary handoff delays incured by horizontal handoffs. Moreover, the throughput measured immediately after a handoff is affected quite considerably by spurious retransmissions due to failed Selective Acknowledgment messages (SACKs) and data retransmission lost. This paper proposes a new Hierarchical Transport layer Mobility protocol (HTM) that deals with local and global mobility and improves throughputs during the handoff period. HTM exploits the dynamic address reconfiguration feature of SCTP and introduces an Anchor Mobility Unit (AMU) in order to complete more efficient handoff procedures. Simulation and numerical results reveal that HTM guarantees lower handoff latency and packet loss, good throughput and limited signaling load compared to mSCTP (mobile SCTP) based mobility

KEYWORDS: Heterogeneous networks, mobility management, SCTP, end-to-end roaming.

3.1 Introduction

The next generation of mobile communication systems, referred to as 4G, 3G+ or beyond 3G, is intended to integrate both current and emerging mobile networks around an IP backbone. For example, this will include second and third generation cellular networks (2G and 3G), satellite systems, Wireless Local Area Networks (WLANs), amongst others. Since each technology is tailored to reach a particular market or a specific type of user services, integrating these heterogeneous systems becomes highly interesting as they offer many possibilities to increase bandwidth, Internet accessibility and area coverage. For example, a mobile user may choose to access a WLAN to send a large data file, but selects a 3G cellular network to place a voice call. However, implementing this type of integrated system implies numerous challenges in mobile handset design, wireless system discovery, terminal mobility, security and billing (Frattasi et al., 2006). Mobility management remains the most significant task to be investigated since it aims to guarantee mobile users disruption-free connections while roaming through heterogeneous networks. Traditionally, mobility management comprises *location management* and *handoff management* (Akyildiz et al., 1999).

Location management is a process which allows networks to localize mobile users' current attachment point for data delivery. Handover or handoff management enables the network to sustain mobile user connections, while they move and change network access points. Handoff mechanisms are usually categorized into: hard and soft handoffs. A hard handoff, also known as *break-before-make*, is completed by first disconnecting with the current access point before switching to another one. This type of handoff mechanism is particularly suitable for delay-tolerant communications traffic. On the other hand, the soft handoff also known as *make-before-break*, is employed by establishing a connection with a new access point before disconnecting from the existing point of attachment. This category of handoff mechanism is particularly suitable for handling latency-sensitive communication services such as videoconferencing. In this sense, Mobile IP (Perkins, 2002) and its further enhancements such as HMIPv6 (Soliman et al., 2005), FMIPv6 (Koodli, 2005) and FHMIPv6 (Jung et al., 2005) are considered among the IETF standards widely accepted to deal with mobility management. However, this category of mobility schemes suffers from weaknesses such as handoff latency, packet loss and signaling load pertaining to the number of bindings to be executed. In addition, certain mobility schemes based on TCP (Snoeren et al., 2000) and SIP (Handley et al., 1999) have been investigated as alternate solutions to the traditional mobile IP. Generally, these proposals need tremendous modifications in both protocol stacks and network architecture (Wei et al., 2005). With the standardization of SCTP (Stewart, 2007), and more particularly with its novel ADDIP Extensions (Stewart et al., 2007), more attention has been paid to experiment mobility over the transport layer. Actually, the transport layer mobility schemes do not depend on the underlying infrastructures and offers the possibility to control the flow and to pause transmission in expectation of a handoff. Thus, a number of solutions which exploit the multihoming features of SCTP have been introduced. Yet, to the best of our knowledge, none of these proposed approaches deal with local mobility at the transport level. This means that current SCTP-based mobility proposals focus on the multihoming feature and do not consider the fact that most of the MN's handoffs are completed inside the same wireless technology (i,e., horizontal handoff). Note that inside an homogeneous technology, an MN may not simultaneously use its two wireless interfaces for communication (Atallah et al., 2006). Obviously, this leads to superfluous delays due to L2 handoff, movement detection, authentication and address configuration. Moreover, certain hidden effects pertaining to fast handovers, such as failed SACKs (Selective Acknowledgements) are not addressed.

The main concern of this paper is to propose a new Hierarchical Transport layer Mobility scheme (HTM) that takes into account local and global mobility in order to reduce handoff latency, packet loss and signaling costs. Additionally, the problem of spurious retransmissions due to failed SACKs and data retransmission lost is addressed. Finally, several simulations and an analytical model are investigated in order to demonstrate the effectiveness of the proposed mobility scheme. In the rest of this paper, the terms mobile user and mobile node will be used interchangeably.

The remainder of this paper is structured as follows: Section 2 presents related work and Section 3 describes the proposed mobility scheme. An analytical model is introduced in section 4. Performance analyses and simulation results are presented in Section 5. Finally, Section 6 concludes the paper.

3.2 Related work

The IP layer is traditionally considered as the default place where mobility is implemented since the IP protocol remains widely used to connect heterogeneous communication systems. However, an increasing interest is recently given to experience mobility at the transport and application levels. In this section, we give an overview of the well-known mobility mechanisms available in the literature.

3.2.1 IP layer mobility

Traditionally, mobility management is performed at the network layer due to the use of the Internet Protocol (IP) that allows routing packets between different technologies. In this context, several approaches propose coping strategies for IP layer mobility. Among these, Mobile IPv6

(MIPv6) is the most popular mechanism that allows mobile nodes to remain reachable in spite of their movements within IP-based mobile environments. However, MIPv6 has some well-known drawbacks, such as high signaling overhead, packet loss and handoff latency, thereby causing real-time traffic deterioration which can be perceived by users (Pe´rez-Costa et al., 2003). These weaknesses led to the investigation of other solutions designed to enhance MIPv6. The IETF proposed new MIPv6 extensions including Hawaii (Ramjee et al., 2002), Cellular IP (Campbell et al., 1999) and Hierarchical MIPv6 (HMIPv6). These protocols tackle intra-domain or micro-mobility, while MIPv6 is used for inter-domain or macro-mobility. However, this solution generates extensive bidirectional tunneling as long as the mobile moves inside the same administrative domain. Additionally, FMIPv6 was proposed to reduce handoff latency and minimize service disruption during handoffs pertaining to MIPv6 operations, such as movement detections, binding updates and address configurations. Although FMIPv6 paves the way for improving MIPv6 performance in terms of handoff latency, it does not efficiently reduce signaling overhead (due to new messages being introduced and exchanged for handoff anticipation) nor does it prevent packet loss (due to space requirements). This may lead to unacceptable service disruptions for real time applications. Combining HMIPv6 and FMIPv6 motivates the design of Fast Handover for HMIPv6 (FHMIPv6) to increase network bandwidth efficiency. However, FHMIPv6 may inherit drawbacks from both HMIPv6 and FMIPv6, those pertaining to synchronization and signaling overhead issues, for instance. Furthermore, the IETF has also proposed a network-based mobility referred to as Proxy Mobile IPv6 (Gundavelli et al., 2008) to ensure mobile user roaming without its participation in any mobility-related signaling. However, this type of mobility schemes depends entirely on the network infrastructure and need a permanent bidirectional tunnel between the MN and CN.

3.2.2 Application layer mobility

Handling mobility at the application layer has also received a lot of attention since this category of solutions is almost independent of the underlying technologies. To accomplish this type of mobility, the SIP (Handley et al., 1999) protocol is widely used. Thus, when a mobile node moves during an active session into different network, it first receives a new address, and then sends a new session invitation to its correspondent node. Subsequent data packets are forwarded to the MN using this new address. However, SIP by itself does not guarantee the maintenance of established Transmission Control Protocol (TCP) sessions or User Datagram Protocol (UDP) port bindings when moving, so further extensions such as S-SIP (Zhang et al., 2007) are needed to provide seamless handover capabilities.

3.2.2 Transport layer mobility

Recently, transport layer-based mobility is gaining attention since it does not require a concept of home network and mobile nodes can perform smooth handovers if they are equipped with multiple interfaces. Moreover, this category of mobility schemes may benefit from flow control and the possibility to pause transmission during the handoff period. The first transport layer mobility solutions were based on TCP, and then other interesting mobility approaches have been proposed with the standardization of SCTP (Stewart, 2007) and mSCTP (Stewart et al., 2007).

3.2.2.1 TCP-based mobility

In the last few years, several transport layer mobility schemes have been proposed to benefit from the connectivity facilities and flow control offered at the transport level. From this perspective, a new TCP protocol architecture was proposed to support mobility (Hsieh et al., 2003). However, tremendous changes must be performed over the entire network to reach this goal. MSOCKS (Maltz et al., 1998) is another TCP-based proposal which does not require changes to the network layer infrastructure. However, it suffers from high latency and packet loss, since it follows a make-after-break approach (disable MN connections until a new path is ready). Migrate (Snoeren et al., 2000) is another TCP-based mobility solution which aims to ensure transparent TCP connection migration. Nevertheless, this solution requires changes to TCP implementation at both ends of the connection. Multi-homed TCP, introduced by (Huitema, 1995), aims to use several addresses in parallel for the same connection by proposing to use new TCP Protocol Control Bloc (PCB) to name the TCP socket, thereby allowing underlying IP addresses to change. However, this approach needs huge modifications and remains, accordingly, not used.

3.2.2.2 SCTP-based mobility

Performing mobility on the transport layer becomes more realistic with the emergence of the Stream Control Transmission Protocol (SCTP), and even more so with its mobile extension. Indeed, SCTP is a new transport layer protocol that was recently standardized under the RFC 4960. It inherited many TCP properties, but it also introduces novel and interesting features, such as multistreaming and multihoming. Multistreaming consists of delivering independent data streams by decoupling reliable deliveries from message ordering. This feature prevents receiver head-of-line blocking in cases where multiple independent data streams occur during a single SCTP session. On the other hand, multihoming allows an SCTP node to be reached through multiple IP addresses (interfaces). In fact, two SCTP nodes can exchange data by defining a common association. In SCTP terminology, an association is equivalent to a TCP connection. End points can be single-homed or multihomed. When single-homed, SCTP nodes are defined as [IP address: SCTP port], otherwise they are designated as [IP1 address, IP2 address…IPn address: SCTP port]. When establishing an association, end points define their primary path, as well as the secondary ones. The primary path is used to transfer data, while secondary paths are used for retransmissions and backups in the event of primary path failures. The SCTP ADDIP (Stewart et al., 2007) Extension enables SCTP nodes to dynamically add, delete and modify their primary address without terminating an ongoing association.

In (Ma et al., 2004) the authors propose an approach to ensure vertical handoffs between UMTS and WLAN networks using SCTP multi-homing capabilities. In (Fu et al., 2004), a TraSH mobility scheme was proposed to perform seamless handovers between heterogeneous networks. In SIGMA (Fu et al., 2005), the authors propose an SCTP-based mobility architecture that integrates location management to ensure seamless handovers. In (Koh et al., 2004), the authors advance certain triggering rules to improve throughput during SCTP-based handoffs. All of these proposals are based on the mobile SCTP extension (mSCTP) and their corresponding mobility procedure is summarized in Fig. 3.1.

1. ***Obtain an IP address from a new location.***
2. ***Add the new IP address to the association.***
3. ***Change the primary IP address.***
4. ***Delete the previous IP address from the SCTP association.***

Figure 3.1 Mobile SCTP-based handoff procedure

In (Iyengar et al., 2006)(Fracchia et al., 2007), the authors put forward new transmission techniques by attempting to enable SCTP-based mobility schemes with concurrent multi-path data transfers. Unfortunately, all of the proposed schemes focus on the inter-system handoffs (i,e., vertical handovers) and do not consider the fact that the majority of handoffs are performed inside the same wireless system (i,e., horizontal handoffs). Accordingly, mobile users must endure unnecessary handoff delays and signaling loads which may become significant in case of frequent handovers. Moreover, a number of hidden effects such as spurious retransmissions due to failed SACKs considerably reduce throughput during handoff periods.

Besides the aforementioned proposals, the Host Identity Protocol (HIP) is introduced to operate in a new layer between the network and the transport layers. The HIP protocol aims to separate the identity (end points and host identifiers) and location information (IP routing) by introducing a new name-space, the Host Identity (HI). The HI is basically a public cryptographic key of a public-private key-pair. A host possessing the corresponding private key can prove the ownership of the public key, i.e. its identity. The separation of the identity and locator makes it is also simpler and more secure to handle mobility and multi-homing in a host. However, this kind of solution suffers from high overhead for short transactions (handshake), lack of micro-mobility and simultaneous node movement capabilities.

3.3 Hierarchical Transport layer Mobility (HTM)

This section offers a detailed description of the proposed Hierarchical Transport layer Mobility (HTM) that copes with local and global mobility at the transport level and addresses the problem of deterioration of throughput during the handoff period. More specifically, a functional scenario is first introduced. Then, the various elements pertaining to the proposed HTM are presented. Note that security issue is out of the scope of this paper.

3.3.1 Functional Scenario

This subsection presents a functional scenario that aims to outline some critical issues that must be addressed when designing a novel SCTP-based mobility scheme.

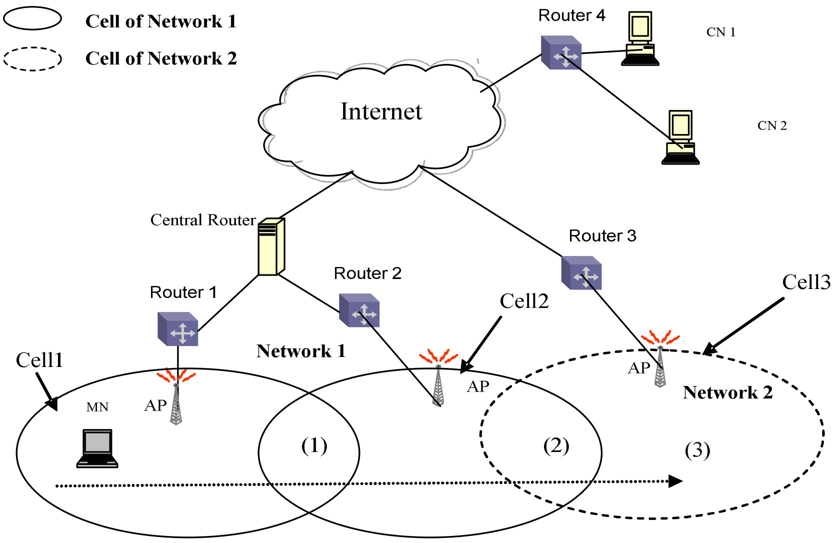


Figure 3.2 Functional scenario

Fig. 3.2 illustrates a very common scenario for an MN that moves through homogeneous/heterogeneous networks. *Network 1* and *Network 2* refer to different mobile technologies (heterogeneous networks), and it is assumed that the MN is multihomed and equipped with two wireless interfaces. The MN, CN1 and CN2 are supposed to support the SCTP protocol.

Initially, the MN has established an association with CN1 and receives its data through *AP 1*. Once the MN enters into the overlapping area (Position (1)), it initiates a horizontal handoff (intra-system) based, for instance, on the quality of the received signal. However, in most radio systems, the MN cannot simultaneously use its two interfaces when it moves inside a same wireless technology. Hence, the delay corresponding to this type of horizontal handoff includes delays relevant to L2 handoff, movement detection, address configuration and association updates. Thus, without taking into account local handoffs, the MN incurs unnecessary handoff delays. Moreover, when an MN changes its primary path, a number of SACKs sent to the MN's previous location are lost as it is shown in Fig. 3.3. Note that the same situation occurs when the CN acts as the sender.

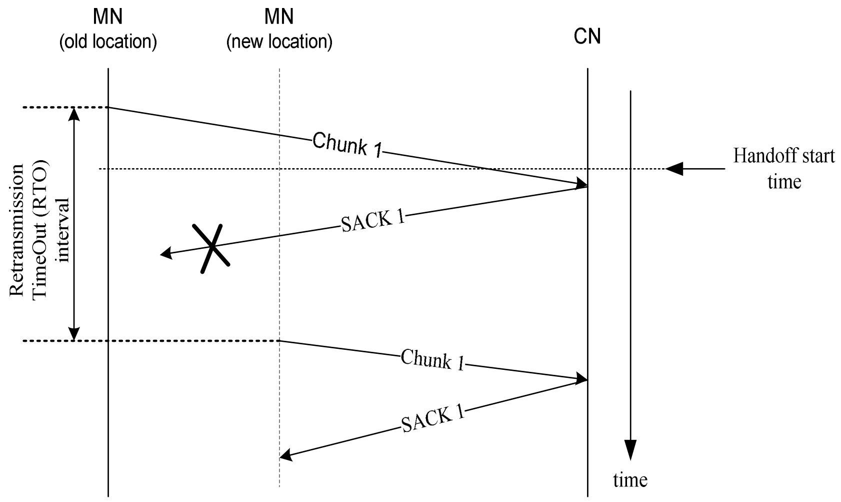


Figure 3.3 Example of failed SACK due to primary path changes

Indeed, the RFC 4960 (Stewart, 2007) states that "an endpoint SHOULD transmit reply chunks (e.g., SACK, HEARBEAT ACK, etc.) to the same destination transport address from which it received the DATA or control chunk to which it is replying; and when its pair is multihomed, the SCTP endpoint SHOULD always try to send the SACK to the same destination address from which the last DATA chunk was received". As a result, a number of SACKs transmitted through a previous path fails to reach their destination since the MN has changed its primary IP address. Consequently, unnecessary Congestion Window (CWND) reductions ensue. Under such circumstances, one may expect that the throughput will be affected. Additionally, when the MN operates as a receiver, a number of data chunks sent to the MN's old primary path will be lost due to a handoff event. Furthermore, all the retransmissions performed after the expiration of the retransmission timeout (RTO) will be also lost as it is shown in Fig. 3.4. Accordingly, a reduction of the CWND parameter will follow. It is clair that such a phenomenon will have a serious impact on the throughput observed during the handoff period.

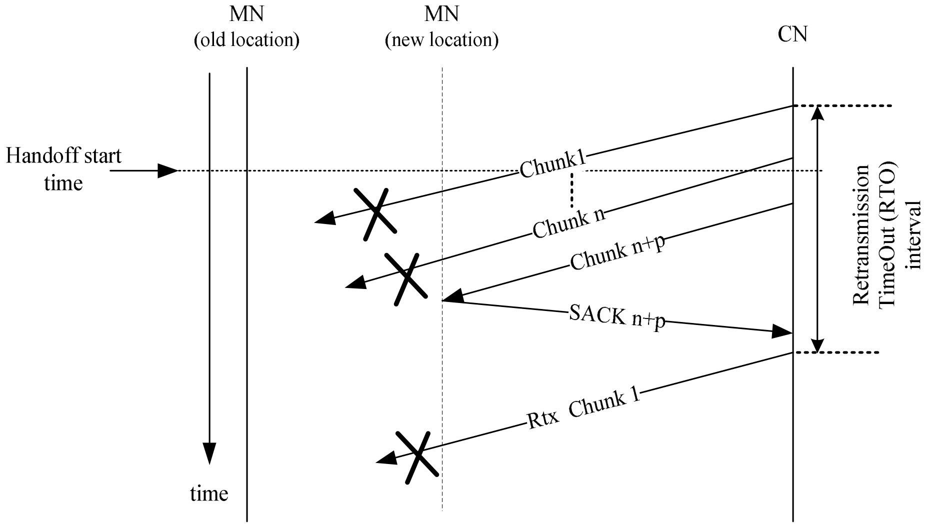


Figure 3.4 Example of failed chunks due to primary path changes

Also, consider the MN located in Position (3) and the CN2 wants to initiate a new SCTP association with it. In the absence of a location management mechanism, CN2 cannot localize its pair (MN). Thus, the MN is prevented from taking advantage of its available wireless interfaces.

The following section introduces our proposed hierarchical mobility mechanism that deals with local and global roaming, and addresses the problem of spurious retransmissions due to failed SACKs and data chunks. Then, a new location management scheme is proposed to ensure the MN tracking.

3.3.2 HTM Architecture

In order to address the aforementioned drawbacks, we propose a novel Hierarchical Transport layer Mobility scheme (HTM) that considers local and global mobility. More specifically, HTM aims to exploit existing hierarchical topologies to implement its new Anchor Mobility Unit (AMU) which allows mobile users to perform local handoffs. In fact, topologies that use hierarchical routers (as illustrated in Fig. 3.1.) are frequently encountered in wireless network designs. Hence, routers (or central routers) that may integrate AMU functionalities can be easily found. Basically, HTM consists of a two-unit handoff procedure designed as: HTM^{local} and HTM^{global}. The former treats local/intra-domain mobility, while the latter deals with global/inter-domain roaming.

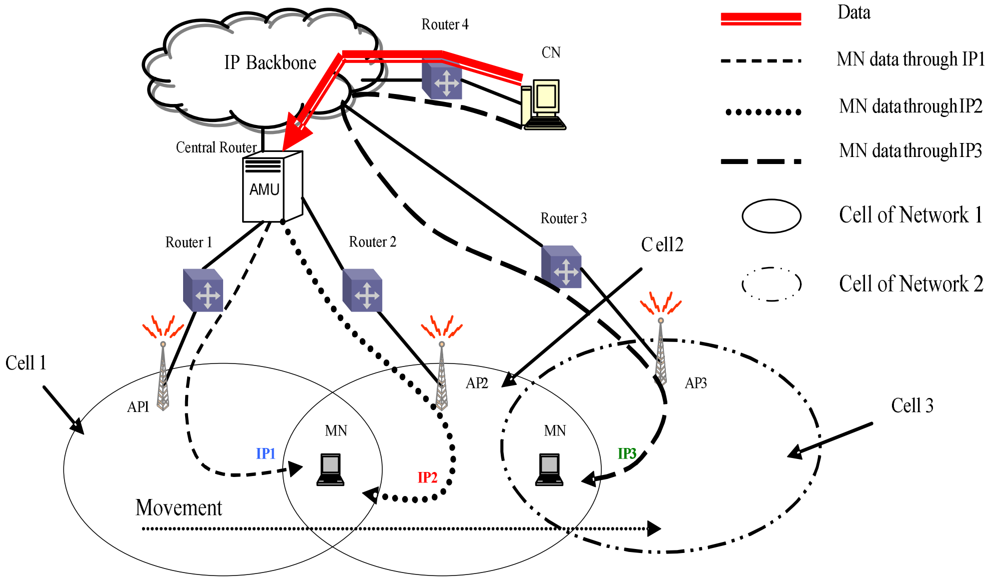


Figure 3.5 HTM architecture

The HTM architecture that supports both local and global handoffs is illustrated in Fig. 3.5 In this architecture, the MN is assumed to be multihomed with two active wireless interfaces. Initially, the MN is assigned to *Cell 1* and receives data from its Correspondent Node (CN) on its IP1 interface. While moving, the MN changes its point of attachment from *Cell 1* to *Cell 2* and finally to *Cell 3*. When the MN hands off from *Cell 1* to *Cell 2*, it performs a local/intra-domain handoff. However, when it moves from *Cell 2* to *Cell 3*, it completes a global/inter-domain handover. Additionally, *AP1* and *AP2* belong to the same wireless system, while *AP3* belongs to an external mobile system. *Router1* and *Router2* are connected to a Central Router (*CR*) which supports AMU functionalities, whose main role is assisting mobile nodes to perform seamless handoffs. Each AMU is identified by an AMU-ID (AMU-Identifier), which is periodically broadcasted in the AP/AR beacons. AMU-IDs are highly useful for MNs to decide whether to perform local or global handoffs. Basically, the AMU functionalities consist of buffering traffic during the disruption period and performing redirection when the MN is attached to the new link. The main AMU process is depicted in Fig. 3.6.

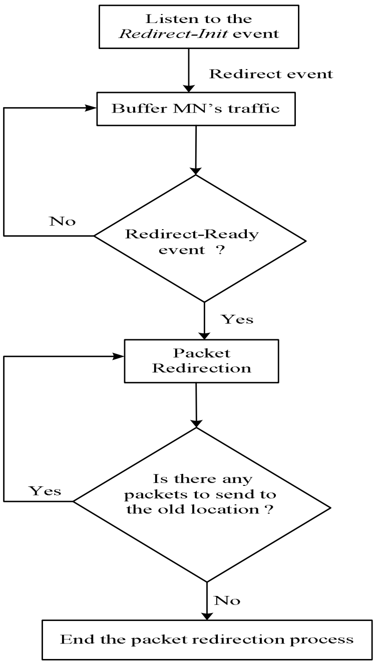

Listen to the *Redirect-Init* event
Redirect event
Buffer MN's traffic
No
Redirect-Ready event ?
Yes
Packet Redirection
Yes
Is there any packets to send to the old location ?
No
End the packet redirection process

Figure 3.6 The AMU redirection process

More specifically, the AMU continuously listens to the redirect events (*Redirect-Init*). Once a *Redirect-Init* event occurs, the AMU starts buffering traffic sent to the old MN's IP address. When the MN is attached to its new location, it sends a *Redirect-Ready* message to notify the AMU that it is ready to receive data on its newly configured IP address. The AMU redirect process ends when no more packets are sent to the old MN address. The following section provides further details pertaining to the proposed handoff procedures when dealing with local and global mobility.

3.3.3 HTM Handoff Procedures

To take benefit of the SCTP multihoming feature we have to remember that when a mobile node moves between cells belonging to a same technology, it can use only one wireless interface a time. However, the MN can simultaneously use its two wireless cards when it moves through cells belonging to heterogeneous technologies. Thus, if we take into account the fact that mobile devices will become increasingly powerful, intelligent and sensitive to link changes, we can assume that the MN detects its movement toward a new access router by using L2 triggers (ie., weak signal strength, high bit error rate, etc.).

As pointed out earlier, the MN detects the presence of the AMU unit through the periodic beacons received from its current point of attachment. Hence, when the MN receives L2 trigger, it sends a *RAS_req* (Router Address Solicitation request) message to its serving AMU to obtain a new address from the next access router (NAR). Accordingly, if the MN receives a new IP address, it concludes that it has to perform an HTM^{local} procedure (local handoff). Otherwise, it runs the HTM^{global} procedure (global handoff).

3.3.3.1 HTM Local Handoff Procedure (HTM^{local})

The HTM^{local} procedure is initiated when an MN perform a handoff, for example from *Cell 1* to *Cell 2*, as illustrated in Fig. 3.5. In this case, it obtains an IP address from *AR2* through its serving AMU unit. Practically, this task can be completed with DHCP (Droms, 1997) or IPv6 autoconfiguration (Thomson et al., 1998). The AMU keeps an association between the new obtained address and the one currently used by the MN. From this moment, the MN is ready to perform a handoff. Recall, that until now the MN continues to receive data from its old path. When the MN decides to move to its new location, it sends a *Redirect-Init* message to the AMU unit. This message informs the AMU that the MN is performing a L2 interface switching (L2 handoff). At this time, the AMU buffers all the packets sent to the MN's previous address until the MN attaches to NAR's link. As soon as the MN is attached to the new access router (NAR), it sends a *Redirect-ready* message to notify the AMU that it has been successfully attached to its new location. Upon receiving the *Redirect-ready* message, the AMU starts packet forwarding to the new MN's IP address. At the same time, the MN sends an *ADDIP_Soft* chunk to inform the CN that a handoff had occurred and it has to set the new MN's IP address as the primary address of their association. Finally, when the MN is completely far from its previous attachment point, the old path is deleted. The entire HTM^{local} procedure is illustrated in Fig. 3.7.

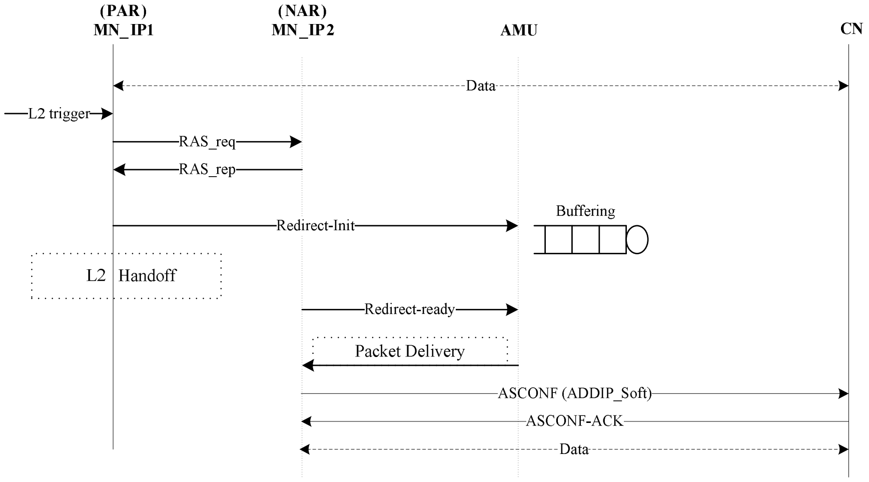


Figure 3.7 The HTM^{local} handoff procedure

ADDIP_Soft is a new chunk introduced to perform the set primary path when the MN is subject to a local handoff. When the CN receives the *ADDIP_Soft* chunk, it concludes that its pair (MN) has performed a local handoff. The CN immediately transmits packets through the MN's new IP address (IP2) and ignores the previous one (IP1). The description of the new proposed *ADDIP_Soft* chunk appears in Fig. 3.8.

<table>
<tr><td>Type = 0xC008</td><td>Length = 20</td></tr>
<tr><td colspan="2">Chunk-ID = 0x11122233</td></tr>
<tr><td colspan="2">Value = 0x0a010101 (New address)</td></tr>
<tr><td colspan="2">Value = 0x0a010111 (Old address)</td></tr>
</table>

Figure 3.8 Description of the *ADDIP_Soft* chunk

The main advantage of the proposed HTM^{local} consists of allowing the MN to perform fast handoffs when an AMU component is available. This strategy adopts a similar principle used in HMIPv6, but the main difference resides in the fact that the tunnel established between the AMU and the MN operates only during the handoff period. The tunnel becomes obsolete when the CN receives the ADDIP-Soft chunk and there is no more packets sent to the old MN's path. This approach is completely different from HMIPv6 and Proxy Mobile IP principles where the tunnel is maintained as long as the MN moves inside the same administrative domain. Additionally, the Network Address Translation (NAT) concept is not suitable in our case since NAT is not designed for mobile purposes. Moreover, many applications and protocols need to use real end-to-end IP addresses. For instance, this is the case with IP security architecture that cannot work across a NAT device since the original headers are digitally signed. The proposed HTM is expected to reduce latency and limit signaling load over the network. Additionally, the problem of spurious retransmissions due to failed SACKs are solved since all messages (including SACKs) destined to the MN are forwarded to the MN through the AMU unit. Finally, note that the AMU unit is implemented over an existing architecture. Hence, in cases where adding an AMU component would be impossible, the MN can perform its handovers by using the HTM^{global} procedure.

3.3.3.2 The HTM Global Handoff Procedure (HTM^{global})

In the absence of an AMU unit, all handoffs are completed with the HTM^{global} procedure described in Fig. 3.9. However, handoffs performed in this case (i.e., without an AMU) may be either horizontal (i.e., same technology) or vertical (i.e., different technology). When the handoff is performed within a same technology (i.e., horizontal handover), the handoff disruption time will include, in this case, L2 handoff movement detection, authentication and address configuration and association update (ADDIP and Set-Primary). However, when the MN performs a vertical handover, the two wireless interfaces can be used simultaneously. Thus, L2 handoff, movement detection, authentication, address configuration and association update (ADDIP), can be completed while the MN continues to receive traffic on its old path. When an MN wants to perform a handoff, for example, from *Cell 2* to *Cell 3* (refer to Fig. 3.4), it listens to the AP3 beacons. Then, it obtains a new IP address from *AR3* (i.e, IP3) to configure its second wireless interface.

The rest of the handoff signaling procedure, in the absence of an AMU unit, is given as follows:

1- The MN sends an ASSCONF (Add IP) message to inform the CN that to add a new IP (MN IP3) address to their association.
2- The CN responds with an ASSCONF-ACK acknowledgement.
3- The MN asks the CN to consider IP3 as its primary address by sending the ASSCONF (Set Primary Address) chunk.
4- The CN sets the new IP address as the MN's primary path and returns an ASSCONF-ACK acknowledgement to the MN.
5- The MN's previous primary address is deleted when the ASCONF (Delete) message is sent to the CN.
6- The CN deletes this address and forwards a confirmation message (ASSCONF-ACK).

The HTM^{global} handover procedure is illustrated in Fig. 3.9.

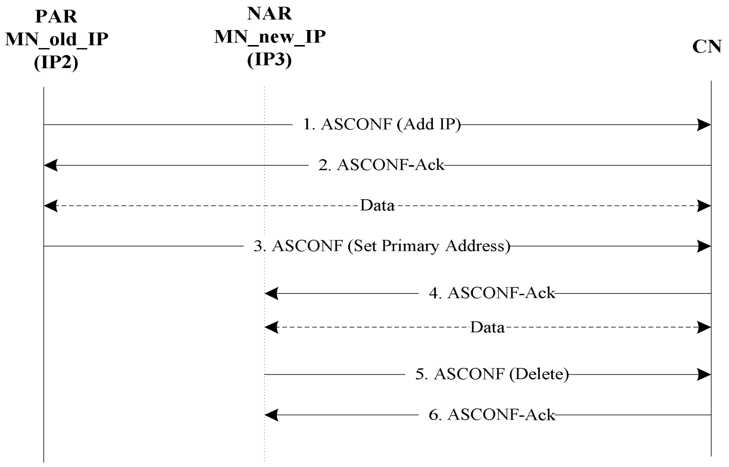


Figure 3.9 HTM^{global} handoff procedure

3.3.4 Analytical model

To study the effectiveness of the proposed HTM, our comparison will consider the mSCTP handoff procedure illustrated in 3.1 since it is, to the best of our knowledge, the only procedure adopted in all the previous mSCTP-based mobility proposals. The conducted analysis focuses on signaling cost, handoff latency and packet loss.

3.3.4.1 Preliminary and notations

Fig. 3.10 illustrates a typical mobility scenario where an MN starts its movement from the X_{start} point and ends at the X_{end} point. During its movement, an MN can perform either handoffs of type (a) or type (b) as indicated in Fig. 3.10. Handoff of type (a) refers to inter AMU domain handover (i.e., local handoff). Handoff of type (b) refers to the end-to-end handover performed outside an AMU domain (i.e., global handoff).

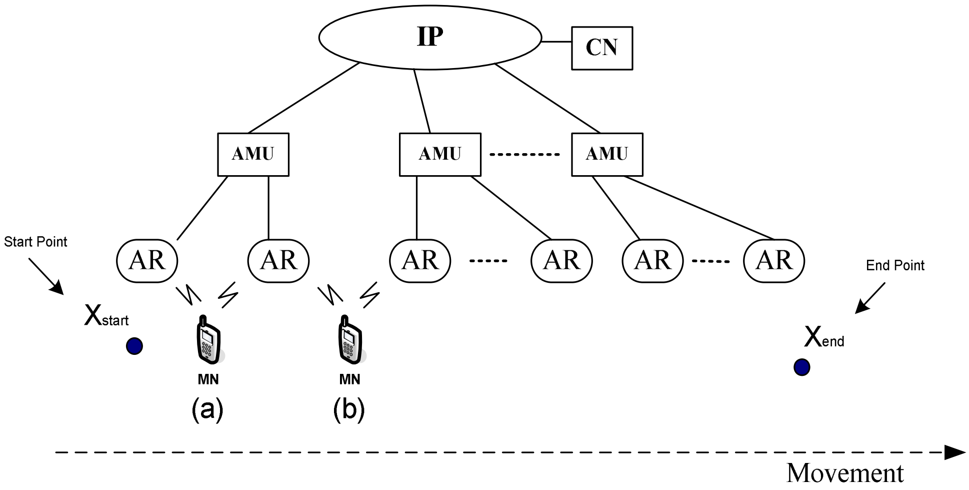


Figure 3.10 MN roaming topology

Let μ_r be the border crossing rate of an MN through access routers (ARs),

Let μ_d be the border crossing rate of an MN through AMU domains,

Let μ_I be the border crossing rate through ARs when the MN remains inside an AMU domain, μ_I is defined as: $\mu_I = \mu_r - \mu_d$. According to (Baumann et al., 1994), if we assume that an AMU coverage area is composed of *M* circular access router subnets, the border crossing rates can be expressed as:

$$\begin{cases} \mu_d = \dfrac{\mu_r}{\sqrt{M}} \\ \mu_I = \mu_r \cdot \dfrac{\sqrt{M} - 1}{\sqrt{M}} \end{cases} \qquad (1)$$

Based on the aforementioned work, μ_r can be defined as: $\dfrac{\rho \cdot v \cdot R_s}{\pi}$, where: ρ is the user density, v the MN average velocity and R_s the perimeter of a subnet.

In order to study the effectiveness of the proposed mobility mechanism we consider a traffic model composed of two levels, a session and packet. The MN mobility will be modeled by the cell residence time and a number of random values introduced in (Fang, 2003). Generally, we model the incoming sessions as a Poisson process (i.e., inter-session arrival time are exponentially distributed). According to (Fang, 2003), the inter-session arrival time may not be exponentially distributed. Thus, alternative distribution models such as Hyper-erlang, Gamma and Pareto have been proposed. However, performance analyses show that the exponential approximation remains an acceptable tradeoff between complexity and accuracy (Fang, 2003). Therefore, for simplicity we assume that the MN residence time in an AR subnet and in an AMU domain follow exponential distribution with parameters μ_r and μ_d respectively, while session arrival process follows a Poisson distribution with rate λ_s. Hence, if we denote: $E(N_r)$ as the average number of AR subnet crossing, $E(N_d)$ as the average number of AMU domain crossing and $E(N_I)$ as the average number of AR subnet crossing performed inside an AMU domain, we can define the above averages as introduced in (Xiao et al., 2004) by:

$$E(N_r) = \frac{\mu_r}{\lambda_s} \qquad (2)$$

$$E(N_d) = \frac{\mu_d}{\lambda_s} \qquad (3)$$

$$E(N_I) = \frac{\mu_I}{\lambda_s} \qquad (4)$$

The notation used in our analysis is given in Table 3.1

Table 3.1 Notation

$T_{X,Y}$	transmission cost between node *X* and node *Y*
P_Z	processing cost at node *Z*
$N_{hop}^{X,Y}$	number of hops between node *X* and *Y*
δ	a proportionality constant to illustrate that the transmission cost for wireless hops are superior to those of wired hops
T_{hop}^{c}	transmission cost per hop
l_X	one lookup cost at node X
η_X	packet tunneling cost at node X
$D_{X,Y}$	transmission delay between nodes *X* and *Y*
$D_{tunneling}$	packet tunneling time
P_Z^t	processing time at node *Z*
T_{MD}	Movement Detection delay
T_{AC}	Address Configuration delay
T_{L2}	L2 handoff delay
T_{UF}	AMU Update and packet Forwarding delay

In what follows, we use the above equations to analyze both signaling and packet delivery costs of the studied mobility schemes.

3.3.4.2 Total cost analysis

We define the total cost (C_{total}) as the sum of signaling and packet delivery costs. In other words, C_{total} is given by:

$$C_{total} = C_{signal} + C_{delivery} \quad (5)$$

The signaling cost refers to the amount of signaling traffic while the packet delivery cost refers to the network overhead. The C_{signal} and $C_{delivery}$ are modeled during an inter-session arrival time that refers to the interval time between the arrival of the first packet of a data session and the arrival of the first packet of the next data session (i,e., one session lifetime). Note that signalling cost required for L2 handoff and address configuration are not considered in our analysis since they are the same for the compared protocols.

a) HTM total cost

The HTM total cost is defined as:

$$C_{total}^{HTM} = C_{signal}^{HTM} + C_{delivery}^{HTM} \quad (6)$$

- HTM signaling cost

The HTM signaling cost is incurred when an MN performs either local or global handoffs. This cost is given by:

$$C_{signal}^{HTM} = E(N_I) \cdot C^{AR} + E(N_d) \cdot C^{AMU} \quad (7)$$

Where :

C^{AR} : refers to the signaling cost when an MN performs a handoff of type (a)

C^{AMU} : refers to the signaling cost when an MN performs a handoff of type (b)

Moreover, if we assume that a handoff preparation is always followed by a handoff execution, the expressions relevant to C^{AR} and C^{AMU} are given in Table 3.2.

Table 3.2 Expression of signalling costs

C^{AR}	=	$T_{MN_p,AMU} + T_{MN_n,AMU} + 2 \cdot T_{MN_n,CN} + 2 \cdot P_{AMU} + P_{CN}$
C^{AMU}	=	$3 \cdot T_{MN_p,CN} + 3 \cdot T_{MN_n,CN} + 3 \cdot P_{CN}$

MN_p

MN_p and MN_n refer respectively to the MN's location before and after a handoff. The $T_{X,Y}$ cost can be expressed as:

$$T_{X,Y} = (N_{hop}^{X,Y} - 1 + \delta) \cdot T_{hop}^{c} \quad (8)$$

To illustrate the impact of the MN's mobility and the MN's average session arrival on the HTM signaling cost, we introduce a session-to-mobility factor (SMR) which represents the relative ratio of session arrival rate to the mobility rate.

The SMR factor is expressed by : $SMR = \frac{\lambda_s}{\mu_r}$ (9).

Hence, if we consider equations (1), (4) and (9), the equation (7) becomes:

$$C_{signal}^{HTM} = \frac{1}{SMR\sqrt{M}}\left[(\sqrt{M}-1)C^{AR} + C^{AMU}\right] \quad (10)$$

- HTM packet delivery cost

Let A_p be the average packets sent by the CN during one session lifetime. Based on Fig. 3.11, the MN can perform either handoffs of type (a) or (b). However, only handoffs of type (a) incur a table lookup and an IP tunneling costs at the AMU. Hence the HTM packet delivery cost is given by :

$$C_{delivery}^{HTM} = A_p \cdot T_{MN,CN} + E(N_I)\cdot(l_{AMU} + \eta_{AMU})\cdot A_p^{(a)} \quad (11)$$

Where: $A_p^{(a)}$ refers to the average packet tunneled during handoffs of type (a),

b) mSCTP total cost

The mSCTP total cost is defined as:

$$C_{total}^{mSCTP} = C_{signal}^{mSCTP} + C_{delivery}^{mSCTP} \quad (12)$$

- mSCTP signaling cost

Based on the mSCTP handoff procedure given in Fig. 3.9, the mSCTP signaling cost is given by:

$$C_{signal}^{mSCTP} = (E(N_I) + E(N_d))\cdot(3\cdot T_{MN_p,CN} + 3\cdot T_{MN_n,CN} + 3\cdot P_{CN}) \quad (13)$$

To express equation (13) as a function of the SMR factor, we use equations (1), (4) and (9).

$$C_{signal}^{mSCTP} = \frac{3}{SMR} \cdot (T_{MN_p,CN} + T_{MN_n,CN} + P_{CN}) \quad (14)$$

- mSCTP packet delivery cost

Since the mSCTP handoff procedure did not incur any IP tunneling or table lookup costs, its packet delivery is given by:

$$C_{delivery}^{mSCTP} = A_p \cdot T_{MN,CN} \quad (15)$$

3.3.4.3 Handoff latency and packet loss

The handoff latency is defined as the time elapsed between sending of the last data packet through the old MN's primary address (i.e., old location) and receiving the first data packet on the MN's new primary address (i.e., new location). The packet loss refers to the amount of packets lost during this disruption time.

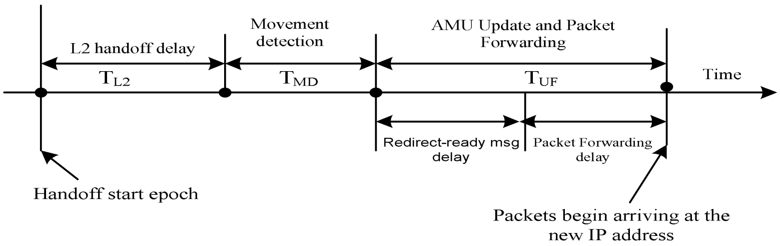


Figure 3.11 HTM local handoff timeline delay

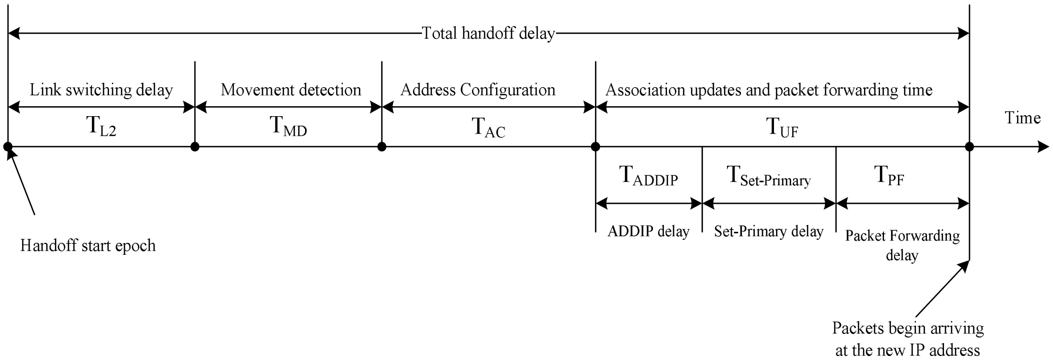


Figure 3.12 mSCTP horizontal handoff timeline delay

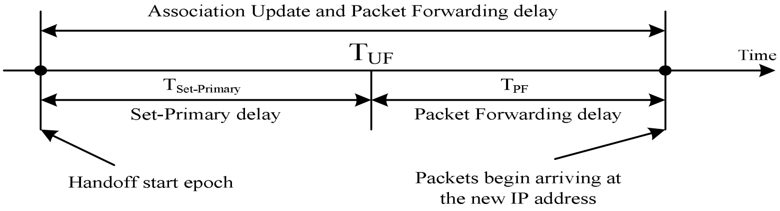


Figure 3.13 mSCTP/HTM vertical handoff timeline delay

If a mobile node moves through cells belonging to a same technology (horizontal handoff), it cannot simultaneously use its two interfaces since it needs two transceivers according to the majority of radio systems (Atallah et al., 2006). However, if it performs a handover between heterogeneous wireless technologies (i,e., vertical handoff), it can use its interfaces in parallel. This means, that the MN continues to receive traffic on its old path while it performs L2 link switching, movement detection, address configuration through the new interface and the association update (ADDIP). Practically, we can divide handoff latency into: link switching or L2 handoff delay (T_{L2}), movement detection delay (T_{MD}), address configuration delay (T_{AC}) and association updates and packet forwarding time (T_{UF}).

According to the handoff scenarios depicted in Fig. 3.10, an MN can perform either handoffs of type (a) or (b). Hence, we define the average handoff latency for HTM as:

$$D_{handoff}^{HTM} = \frac{1}{E(N_I)+E(N_d)} \cdot \left[E(N_I)\cdot D_{handoff}^{(a)} + E(N_d)\cdot (P_h \cdot D_{handoff}^{(b),horizontal} + (1-P_h)\cdot D_{handoff}^{(b),vertical})\right] \quad (16)$$

Where:

$D_{handoff}^{(a)}$: latency relevant to handoff of type (a) (i.e., inside an AMU domain), the corresponding timeline delay is given in Fig. 3.11.

$D_{handoff}^{(b),horizontal}$: latency relevant to horizontal handoff performed outside an AMU, the corresponding timeline delay is given in Fig. 3.12.

$D_{handoff}^{(b),vertical}$: latency relevant to a vertical handoff, the corresponding timeline delay is given in Fig. 3.13.

P_h : probability that an MN perform a horizontal handoff outside an AMU domain.

The expressions of $D_{handoff}^{(a)}$, $D_{handoff}^{(b),horizontal}$ and $D_{handoff}^{(b),vertical}$ are given in Table 3.3.

Table 3.3 Expression of HTM handoff delays

$D_{handoff}^{(a)}$	=	$T_{L2}+T_{MD}+2\cdot D_{MN,AMU}+D_{tunneling}+P_{AMU}^{t}+\tau$
$D_{handoff}^{(b),horizontal}$	=	$T_{L2}+T_{MD}+T_{AC}+4\cdot D_{MN,CN}+P_{CN}^{t}+\tau$
$D_{handoff}^{(b),vertical}$	=	$2\cdot D_{MN,CN}+P_{CN}^{t}+\tau$

If we consider that $\mu_d \succ 0$ (i,e., we have at least two AMU domains), we use equations (3) and (4) to derive the following relation:

$$D_{handoff}^{HTM}=\frac{1}{\sqrt{M}}\cdot\left[(\sqrt{M}-1)\cdot D_{handoff}^{(a)}+P_h\cdot D_{handoff}^{(b),horizontal}+(1-P_h)\cdot D_{handoff}^{(b),vertical}\right] \qquad (17)$$

Where: τ refers to the time between the instant when the sender is ready to send data packets and the instant when it effectively starts sending data packets to the MN's new location.

According to (McNair et al., 2001) $D_{X,Y}$ is defined as:

$$D_{X,Y}=\frac{1-q}{1+q}\cdot(\frac{s}{B_{wl}}+L_{wl})+(N_{hop}^{X,Y}-1)\cdot(\frac{s}{B_w}+L_w+\varpi_q) \qquad (18)$$

Where *s* is the message size, ϖ_q is the average queuing delay at each intermediate router, q is the probability of wireless link failure, B_{wl} (resp B_w) the bandwidth of wireless (resp wired) link and L_{wl} (resp L_w) wireless (resp wired) link delay.

With mSCTP, the handoff latency is given by:

$$D_{handoff}^{mSCTP}=\frac{1}{\sqrt{M}}\left[\left(\sqrt{M}-1+P_h\right)\cdot D_{handoff}^{b,horizontal}+(1-P_h)\cdot D_{handoff}^{b,vertical}\right] \quad (19)$$

On the other hand packet loss is proportional to the handoff delay since all data packets exchanged during this disruption period are lost. Practically, let λ_p be the packet arrival rate, the packet loss for both HTM and mSCTP is defined as:

$$\begin{cases} P_{loss}^{HTM} = \lambda_p \cdot D_{handoff}^{HTM} - Min(B_{HTM}, B_{AMU}) \\ P_{loss}^{mSCTP} = \lambda_p \cdot D_{handoff}^{mSCTP} \end{cases} \quad (20)$$

Where, B_{HTM} is the buffer size required for HTM and B_{AMU} the buffer size available at the AMU. The buffer size required for HTM is proportional to packet arrival rate and it is computed as follows:

$$B_{HTM} = \lambda_p \cdot (T_{L2} + T_{MD} + T_{UF}) \quad (21)$$

3.4 Performance Evaluation

This section presents simulation and numerical results obtained when an MN uses either the proposed HTM or the mSCTP based handoff procedure. We choose mSCTP as the benchmark transport layer mobility protocol for our comparison since all the previous SCTP-based mobility proposals use the mSCTP standard. Moreover, mSCTP is a general IETF purpose standardized under the RFC 5061.

3.4.1 Simulation Setup

The main concern of our simulations is to show how the introduced AMU unit improves handoff seamlessness. That is why we consider the simulation scenario depicted in Fig. 3.14. This scenario is designed in such a way to provide realistic results, while remaining sufficiently small to be handled efficiently with the ns-2 simulator. Simulation code is based on the SCTP module developed at the University of Delaware. This SCTP module is modified so that it can support the newly introduced ADDIP-Soft Chunks, as well as AMU functionalities (Section 3.3.2).

The MN is supposed to be multihomed and equipped with two 802.11b interfaces. Initially, the MN is assigned to AR1 and benefits from an ongoing association with CN. When the MN moves from AR1 to AR2, it performs a local handoff (inside an AMU). In all simulations, the observed MN moves at various speeds, on a straight line, between AR1 and AR2 sub-networks. Each AR operates according to the 802.11b (11 Mbit/s) standards in the Distributed Coordination Function (DCF). Delays for both 802.11b WLANs equal 15 ms. For each simulation, a CBR agent is attached to a CN, as is a sink agent to the MN. The average experiment time lasts around 300 s. In one experiment, the MN can complete several rounds (starts from AR1 to AR2 and returns back to AR1).

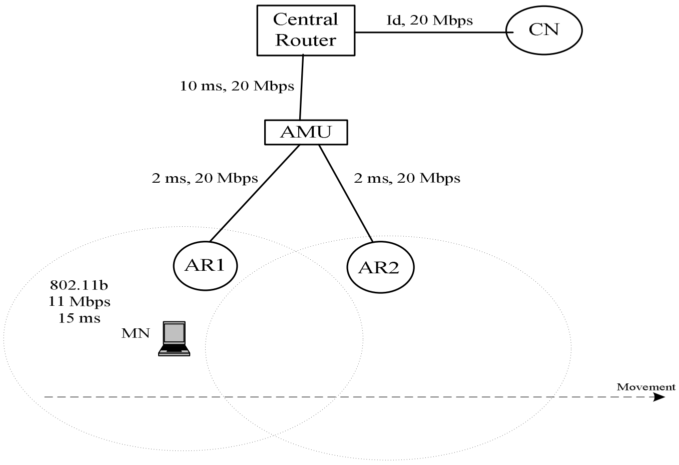


Figure 3.14 Simulation network topology

3.4.2 Simulation Results

Fig. 3.15 illustrates handoff latency behavior when an MN completes HTM^{local} and mSCTP handoffs. In fact, several experiments were conducted where the MN performs a handoff from

AR1 to AR2, then it returns back to AR1. Every time this experiment is performed, a wired hop is added between the MN and the CN, meaning that an additional delay is added to the CN-AMU link. The first thing to be noted is that when the number of intermediate hops between the MN and the CN increases, the mSCTP latency values continue to increase, while HTM^{local} latency remains approximately constant. This situation is due to the fact that HTM^{local} uses the AMU unit to redirect packets to the MN's new location as quick as possible. Then, it updates its association. This approach is completely different from mSCTP that has to update the MN's active association with ADDIP and Set-Primary chunks during the disruption time. Moreover, the HTM^{local} handoff latency remains lower than mSCTP one even if the distance between MN and CN is low. Indeed, with HTM^{local}, the MN anticipates its address configuration process by using the AMU unit (which is not possible with mSCTP). Recall that the address configuration delay may take over than 500 ms (Mishra et al., 2003).

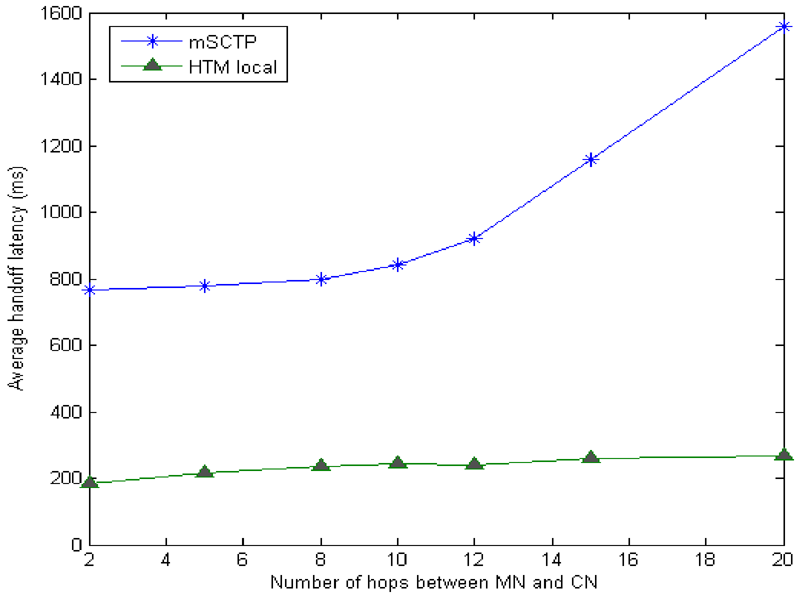


Figure 3.15 Impact of MN-CN distance on handoff latency

Fig. 3.16 presents the average handoff latency, for both mSCTP and HTM, as a function of moving speed. Here, we set id = 20ms (i,e., delay between central router and CN) and we increase the MN's speed (*v*) from 2 m/s to 40 m/s while it performs several handoffs between *AR1* and *AR2*. We notice that when the MN's speed is small, HTM shows lower handoff delay than mSCTP. However, when *v* > 12 m/s, the HTM's latency increases and becomes equivalent to the mSCTP one. This is because when the moving speed increases, the sojourn time in the overlapping area becomes too small, so the MN do not have enough time to perform its configuration process. Moreover, with high values of MN's speed, the handoff delay increase for both HTM and mSCTP since they do not have sufficient time to complete their respective handoff procedures.

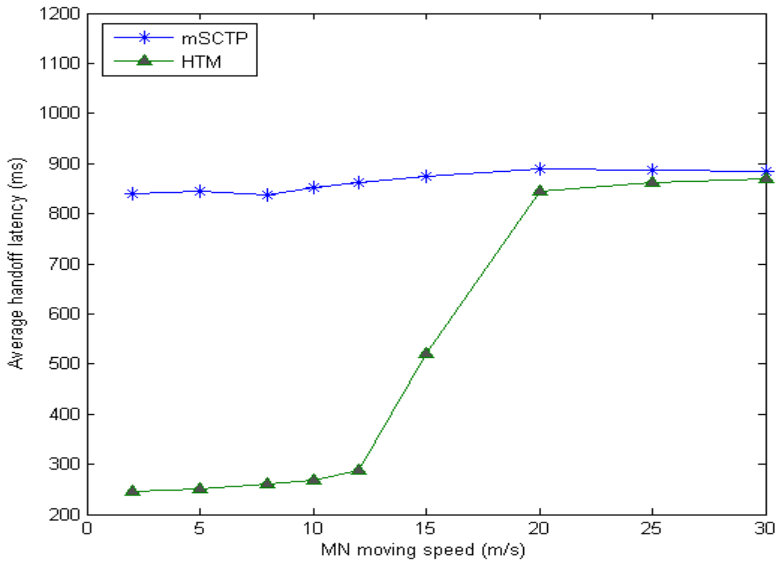


Figure 3.16 Impact of moving speed on HTM / mSCTP latencies

To illustrate how the proposed HTM improves throughput, we will first consider the results illustrated in Fig. 3.17. These results correspond to the throughput relevant, respectively, to the previous and new MN's paths, i.e, the MN changes its point of attachment from AR1 to AR2.

Observe that the throughput of the previous path decreases during the time interval $t \in [13s, 25s]$ where the handoff takes place. This drop is due to the increasing loss rate of AR1 as the MN moves. Once the handoff is over, notice that the MN throughput increases again until it reaches its original level. However, the throughput reported immediately after the handoff remains lower than the one computed before the handoff occurred.

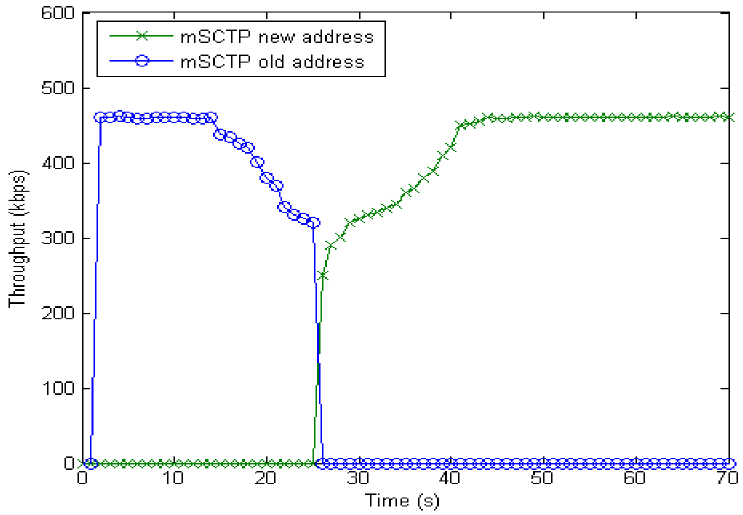


Figure 3.17 Throughput relevant to an mSCTP path handover

This situation is due to failed SACKs that cause a diminution of the congestion window (CWND), thus reducing throughput. To show how the proposed HTM improves throughput compared to mSCTP, consider the throughput obtained immediately after a handoff for HTM and mSCTP.

Fig. 3.18 shows the throughput pertaining to the time interval (25-30s) following an MN handoff. Note that the HTM throughput is relatively high compared that of an mSCTP. This is due to the fact that HTM^{local} uses the AMU unit to buffer and forward all the traffic to the new MN's location. This traffic obviously includes SACKs which are not lost, unlike what happens with mSCTP.

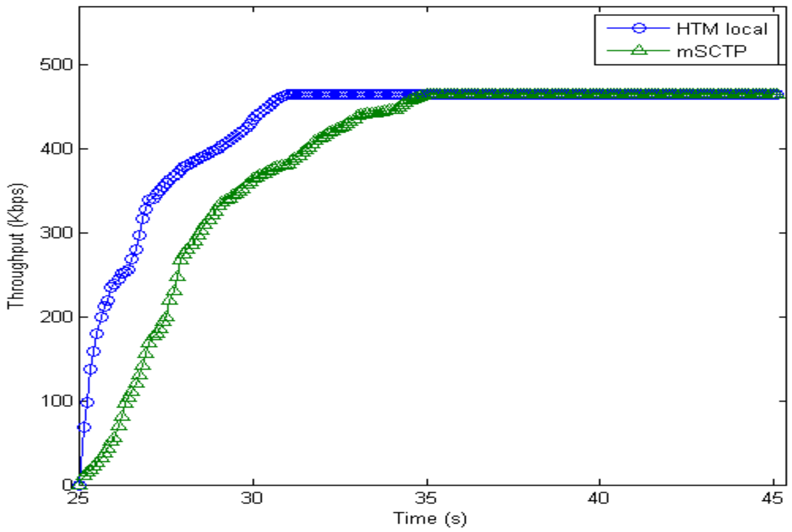


Figure 3.18 Throughput of HTM^{local} vs mSCTP

3.4.3 Numerical Results

In this section, we use the developed cost models (section III) to illustrate how the proposed mobility scheme HTM improves QoS parameters in terms of signalling cost, handoff delay and packet loss compared to mSCTP.

The list of the parameter values used for our numerical results is shown in Table 3.4.

Table 3.4 Parameters used for performance analysis

Parameters	Symbols	Values
Wireless link failure probability	q	0.5
Average queuing delay	ϖ_q	0.1 ms
Wired link delay	B_w	100 Mbps
Wireless link bandwidth	B_{wl}	11 Mbps
Message size	s	296 bytes
Number of AR subnets per AMU/MAP domain	M	4
Average packet arrival per session	A_p	20
Average packets tunneled during a handoff of type (a)	$A_p^{(a)}$	2
Lookup cost at the AMU	l_{AMU}	2
Packet tunneling cost at the AMU	η_{AMU}	2
L2 handoff delay	T_{L2}	50 ms
Movement detection delay	T_{MD}	100 ms
Address Configuration delay	T_{AC}	500 ms
Waiting time before effective data transmission	τ	1 ms

Fig. 3.19 illustrates the total signaling cost as a function of the SMR ratio. When the SMR ratio is inferior to 1, the mobility rate is higher than the session arrival rate that is why the signaling cost increases for both HTM and mSCTP. This increase becomes more noticeable when the SMR is close to 0. However, the HTM cost remains lower than the mSCTP cost. On the other hand when the SMR is superior to 1, i,e., the session arrival rate is greater than the mobility rate, the binding updates, relevant to handoffs, are performed less often.

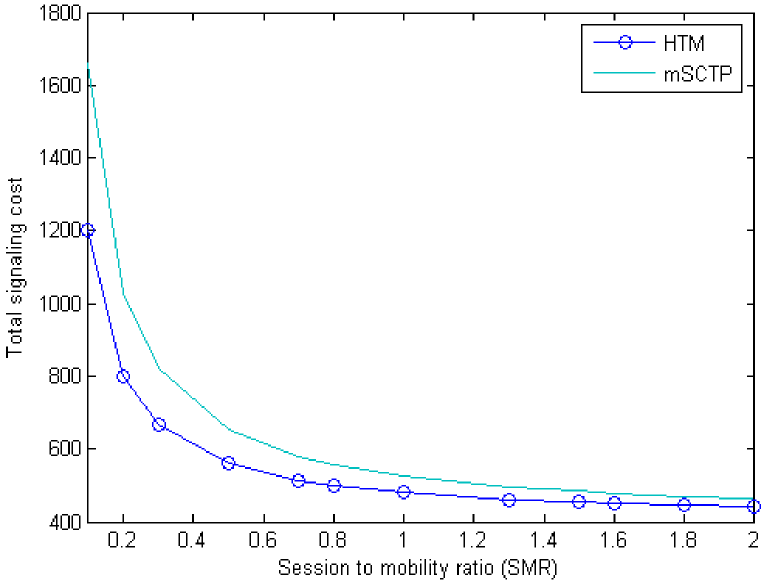


Figure 3.19 Impact of the SMR on the total signaling cost

Fig. 3.20 illustrates the total signaling cost as a function of mobile node velocity. We notice that the total signaling cost increases for both HTM and mSCTP. However, the signaling costs involved by HTM remain lower than mSCTP. Moreover the gap between mSCTP and HTM signaling costs becomes more important when the MN's velocity increases. This behaviour is to be expected since the MN will perform frequent handoffs when its velocity reaches high values. Nevertheless, HTM takes into account local handoffs, hence its relevant signaling costs are lower than mSCTP.

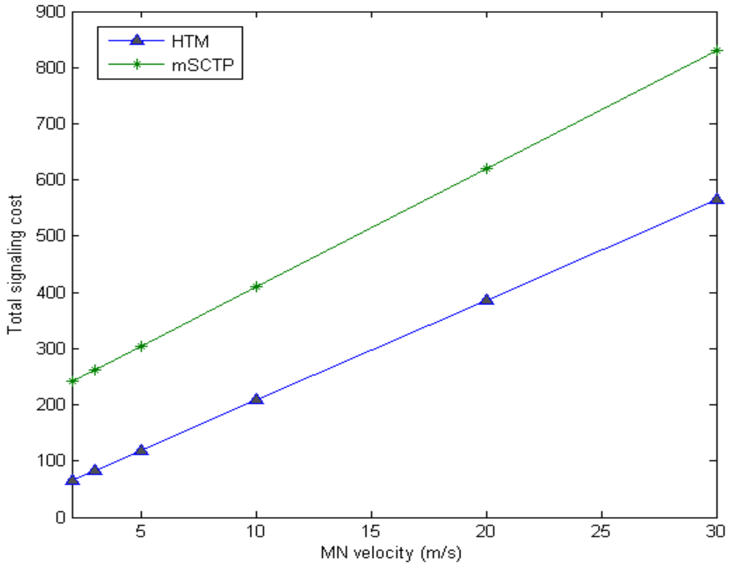


Figure 3.20 Impact of the MN velocity on the total signaling cost

Fig. 3.21 shows that the HTM total signaling cost is proportional to the AMU tunneling cost. However, it remains lower than the mSCTP cost even if high values are used for the AMU tunneling cost (i., more that 20). Recall that all of the processing costs used for our performance analysis are less or equal to 2. On the other hand, mSCTP is not affected by the AMU cost variation since it does not perform traffic redirection.

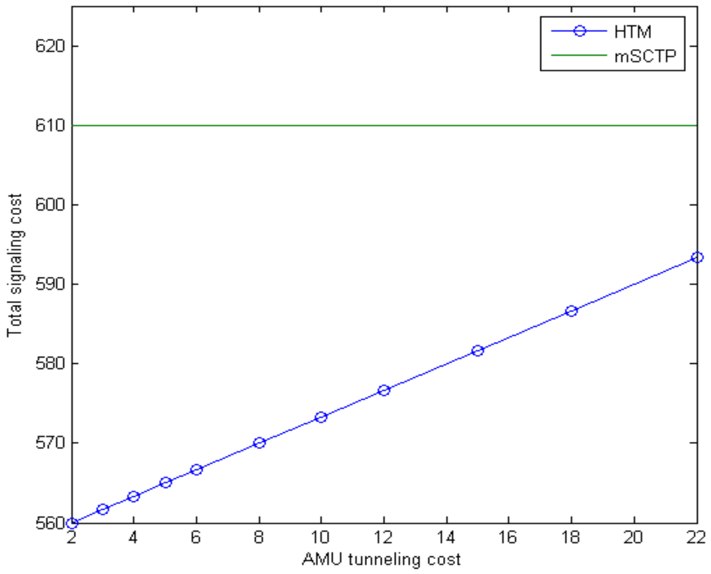


Figure 3.21 Impact of the AMU tunneling cost on the total signaling cost

In Fig. 3.22 we present the average handoff delay as a function of the wireless link delay. We notice that the average handoff delay is proportional to the wireless link delay for both HTM and mSCTP. However, it can be noticed that the HTM average latency is lower than mSCTP. Moreover, when P_h increases (i,e., probability of horizontal handoff performed outside an AMU unit), handoff latencies increase for both HTM and mSCTP. However, the HTM's latency remains lower than the mSCTP one. This means that the introduction of the AMU unit improves considerably the MN's handoff delays during its roaming through homogeneous networks.

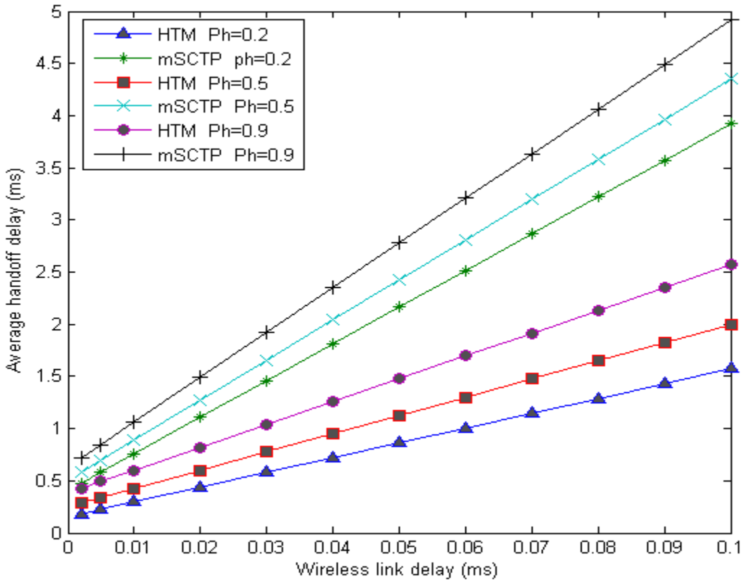


Figure 3.22 Handoff latency as a function of wireless link delay

The impact of HTM^{local} on the MN's latency is clearly illustrated in Fig. 3.23 where we compare the two scenarios of mSCTP handoffs (i.e., horizontal and vertical) with to our proposed mobility scheme. Recall, that HTM and mSCTP use the same vertical handoff procedure.

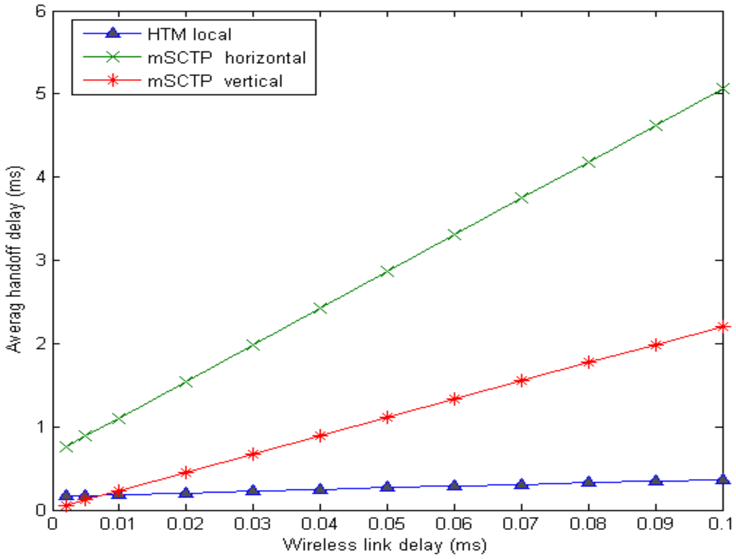


Figure 3.23 Impact of wireless link delay on horizontal/vertical handoffs

Fig. 3.24 illustrates the impact of the AMU tunneling delay on the average handoff latency. We notice that the HTM handoff latency remains lower than mSCTP even if for high tunnelling delays. On the other hand we notice that mSCTP is not affected by the AMU tunneling delay since it does not use the tunneling process while performing handovers.

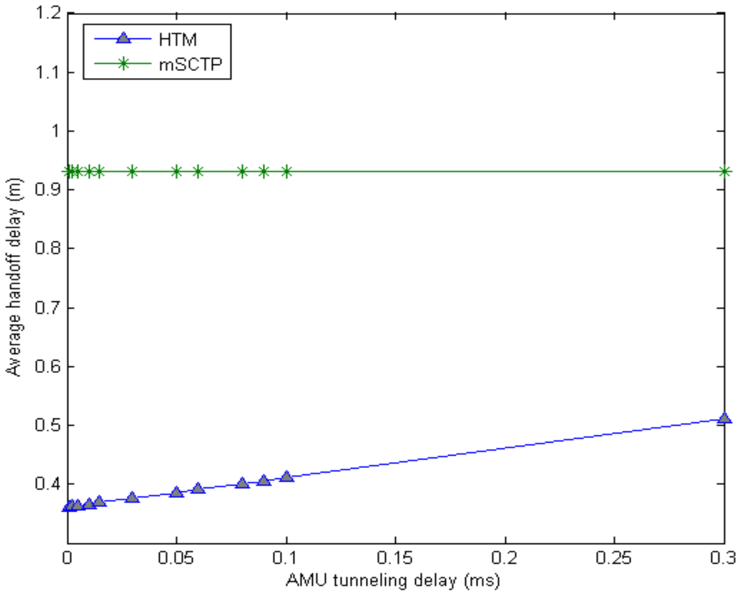


Figure 3.24 Handoff latency as a function of the tunneling delay

Fig. 3.25 illustrates handoff latency as a function of the average subnet crossing rate inside an AMU (*E*(N_I)). When this rate is low, i,e., most of the MN's handovers are performed in the absence of the AMU units, we notice that the average HTM latency is high. With the increase of *E*(N_I), we observe a noticeable decrease of the HTM average latency which becomes approximately constant when this rate reach high values. This situation shows again that the consideration of local handoffs by our mobility proposal reduces considerably the overall average handoff delay when the MN performs consecutive horizontal and vertical handoffs. On the other hand, the mSCTP handoff latency remains high and insensitive to the *E*(N_I) rate.

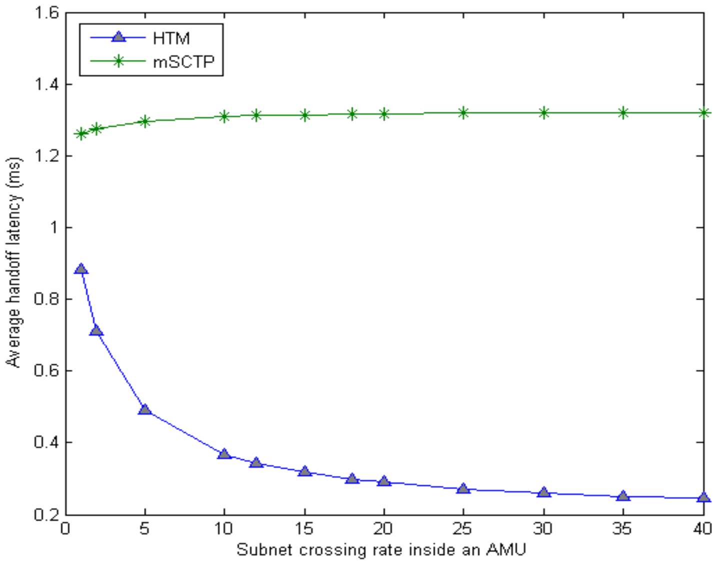


Figure 3.25 Impact of the intra-subnet crossing rate on the handoff delay

Fig. 3.26 shows the behavior of packet loss as a function of packet arrival rate. It is noticed that packet loss increases for both HTM and mSCTP. However, the HTM packet loss remains lower than mSCTP. This situation is quite normal since the handoff delays for HTM is lower than mSCTP and by definition of all of the packets received at this period are lost. In addition, HTM uses a buffering strategy when an MN roams inside a same AMU domain which helps to avoid as much as possible packet loss during the MN's disruption time.

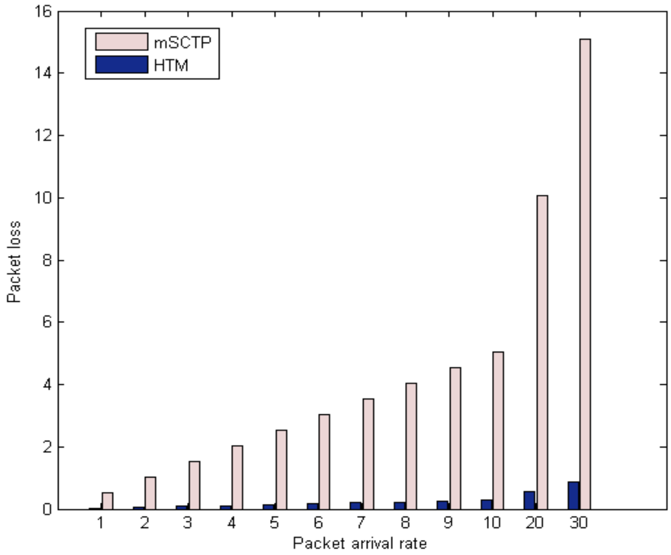


Figure 3. 26 Packet loss behavior for different packet arrival rates

3.5 Conclusion

This paper proposes a new hierarchical transport layer mobility scheme called HTM, whose main goal is to provide mobile nodes with seamless roaming through heterogeneous networks. More specifically, HTM consists of an end-to-end mobility protocol based on SCTP features, which includes multihoming and ADDIP Extension. It particularly introduces an Anchor Mobility Unit (AMU) to deal with local mobility in order to reduce handoff latency and signaling load. Additionally, HTM addresses the problem of spurious retransmissions due to failed SACKs. Simulations and numerical results show that HTM ensures low latency, good throughput and limited signaling load compared to the mSCTP based handoffs. Future work shall investigate how this proposal can be adapted to mobile ad hoc networks as well as the impact of location management on system performance.

CHAPITRE 4

ADAPTIVE DECISION MAKING STRATEGY FOR HANDOFF TRIGGERING AND NETWORK SELECTION

Abdellatif Ezzouhairi, Alejandro Quintero, Samuel Pierre

Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
C.P. 6079, succ. Centre-Ville, Montreal, Quebec, H3C 3A7, Canada
Phone: 514 340-3240 ext. 4685. Fax: 514 340-3240
Email: {Abdellatif.Ezzouhairi; Alejandro.Quintero; Samuel.Pierre}@polymtl.ca

Abstract

Next generation mobile networks are expected to integrate a large number of wireless technologies. However, this integration yields many challenges such as those pertaining to handoff triggering and decision making. Various approaches have been proposed to solve these problems, yet handoff initiation and network selection remain critical issues which are widely based on RSS (Received Signal Strength) measurements. Moreover, the use of context-awareness is very limited in the previous works. This paper proposes a new handoff decision strategy which aims to efficiently deal with handoff triggering and network destination selection with respect to mobile terminal requirements and network capabilities. Furthermore, we introduce a new score function that estimates network preferences for both voluntary and forced handoffs. Additionally, to render easier the accessibility to context information, we develop a context aware mechanism which is based on a third party architecture. Finally, simulation results show that compared to

RSS-based approaches, the proposed handoff decision strategy has greater respect for users' requirements and preferences.

KEYWORDS : Handoff strategy, Decision making, Context-awareness, Handoff triggering.

4.1 Introduction

Rapid progress in wireless network and communication technologies has created a wide variety of mobile systems. For example, Bluetooth is used in indoor areas, IEEE 802.11 in local areas, Universal Mobile Telecommunication System (UMTS) in expanded areas and satellite networks for global coverage. In order to take advantage of these complementary technologies, fourth generation (4G) systems are expected to integrate a large number of these heterogeneous wireless systems (Nasser et al., 2006). According to emergent trends in mobile communication, 4G systems will guarantee seamless roaming and quicker handoffs through heterogeneous technologies. However, seamless mobility is more complex, as integrated environments will support different wireless technologies. The literature commonly refers to such an issue as mobility management (Akyildiz et al., 1999).

Mobility management comprises two components: location management and handoff management. Location management enables the network to track the locations of mobile users between consecutive communications, while handover, or handoff management, refers to the process of transferring a mobile user between cells of the same or a different network without disrupting connections. Handoffs performed between cells that belong to the same network are considered *homogeneous* and the handover is called *horizontal*. This kind of handoff is mainly caused by the movement of the mobile user out of the coverage area of its current cell. On the other hand, handoffs performed between cells that belong to different networks are considered *heterogeneous* and this type of handoff is referred to as *inter-system* or *vertical handover* (VHO) (McNair et al., 2004).

In heterogeneous wireless environments, mobile users can perform both horizontal and vertical handoffs. Horizontal handoffs are similar to those performed in homogeneous systems. However, vertical handoffs are performed between systems built on different wireless technologies. In addition, depending on initiation reasons, handovers can belong to either of two categories: forced and voluntary (Tansu et al., 2006). Handovers caused by low link quality (weak RSS, low bandwidth, high traffic, etc.) are qualified as *forced,* since the mobile node must select a new destination and execute the handoff process very quickly, while the voluntary handoff aims to maximize users' satisfaction. Actually, handoff triggering remains an important issue to be investigated for the next generation of mobile networks since the received signal strength (RSS) measurements are not enough to decide when to initiate handoffs (Kassar et al., 2008). Indeed, an MN may have a good RSS signal but a very low bandwidth or high traffic conditions. It is obvious that, in such circumstances, the MN has to trigger a handoff especially if it carries on a multimedia traffic. Moreover, during handoff triggering an MN must decide whether to trigger a forced or a voluntary handoff since the former aims to avoid QoS deterioration while the latter is used to improve MN preferences.

Once a handoff initiation takes place, the MN has to select its future network destination. This issue is known in the literature as handoff decision or network selection. It consists in selecting a new network destination (or a new access point in case of homogenous networks) that provides best QoS conditions with respect to MN requirements and network capabilities (Siddiqui et al., 2006). The handoff decision is generally driven by metrics which are strictly related to the RSS level and resources availability. However, in 4G, the RSS from different networks do not have the same meaning since each network is composed of its specific characteristics and there is no common pilot signal. Then, RSS comparisons are insufficient for handoff decision and may be inefficient or impractical. A more complex decision criterion that combines a large number of parameters such as monetary cost, bandwidth, power consumption and user preferences is necessary.

Furthermore, handoff decision is typically based on a score function to complete network selection. Thus, the quality of the selected network destination depends on the way the score function is designed. We advocate that an efficient score function must consider both the handoff type (forced or voluntary) and network stability. The handoff type can be used to choose adequate context parameters to conduct network selection while network stability can be considered to eliminate networks that present rapid and high QoS variations. Finally, context awareness is also an important task to be addressed in order to specify how context information will be provided for both handoff triggering and network selection.

Based on the aforementioned motivations, this paper proposes a new handoff decision strategy that deals with handoff triggering and network selection. More specifically, the main contributions of this paper consist in: (1) proposing a handoff triggering scheme based on fuzzy logic to decide which type of handoffs to initiate (forced or voluntary) and under which conditions, (2) designing a handoff preference function that models both forced and voluntary handoffs in order to perform best network selection, (3) proposing a context aware mechanism that ensures data sharing and provide various context information, (4) Analyzing the performances of the proposed handoff decision strategy.

The remainder of this paper is organized as follows: Section II introduces the related work. Section III describes the proposed context aware mechanism. Section IV outlines the proposed handoff decision strategy. Section V presents and discusses the obtained results and finally, Section VI concludes the paper.

4.2 Related work

In the traditional cellular systems, such as the global system for mobile communication (GSM), a threshold comparison of several specific metrics is used to make handover decisions. The most common metrics are Received Signal Strength (RSS), Signal-to-Interference Ratio (SIR) and Bit Error Rate (BER). However, RSS comparisons fail to consider network capabilities and mobile users' options (Jha et al., 2004). Therefore, RSS measurements alone are insufficient for handoff decisions. To overcome this drawback, several handover decision strategies have been proposed in the literature. These proposals can be divided into: multi-criteria, Fuzzy Logic (FL) and Neural Network (NN) based, context-aware, user-centric and decision function based strategies.

First, the multi attribute decision strategies aim to deal with network destination selection among a limited number of candidate networks belonging to different technologies with respect to various criteria. This is known in the literature as multi attribute decision making problem (MADM) (Hwang et al., 1981). The popular MADM resolution methods are: SWA (Simple Additive Weighting), TOPSIS (Technique for Ordered Preference by Similarity to ideal Solution), AHP (Analytical Hierarchy Process) and GRA (Grey Relational Analysis). In this sense, a network selection mechanism that combines AHP and GRA has been proposed in (Quiqyang et al., 2005) to find a tradeoff between user preferences, service application and network conditions. The results revealed that this selection approach can work efficiently for an UMTS/WLAN system. However, MADM based solutions remain insufficient to handle decision with imprecise criteria.

Second, to overcome the weakness of using imprecise parameters in the MADM strategies, Fuzzy Logic (FL) and Neural Networks (NN) concepts are then introduced for network selection.

Hence, an advanced neural-network-based vertical handoff algorithm was developed in (Pahlavan et al., 2000) to satisfy users' bandwidth requirements. However, this type of algorithm is not easy to handle especially with mobile nodes having limited computing and storing capabilities. Additionally, training of the neural network has to be done beforehand. In (Chi-Hsing et al., 1999), the authors proposed a solution incorporating Fuzzy Logic in which terrestrial and satellite mobile networks operate alongside each other. In this case, handover decision aims to select a segment or a network for a particular service that can satisfy objectives based on criteria such as: low cost, good RSS, optimum bandwidth, low network latency, high reliability, long life battery and preferred access network. In (Makela et al., 2000) the FL and NN concepts are used together to provide handoff decision making. However, these solutions lack in using efficient context awareness since networks and operators are very reticent to share their own context information.

Third, the context-aware based handover concept uses context information of both mobile node and networks to take decision whether the handover is necessary on the access network target (Jung et al., 2005). In (Balasubramaniam et al., 2004), the authors present a framework with an analytical context categorization and a detailed handover decision algorithm. Prototype experiments have used different type access networks and streaming applications. It has shown that this approach can be used to deal with handoff selection. However, context information gathering is performed by a single point (context repository) which can cause failure point. Moreover, it needs frequent communication between the MN and the network, resulting in increased overhead on the radio link.

Fourth, user-centric strategies focus on user satisfaction in terms of monetary cost and QoS. More specifically, this type of solutions, propose handover decision policies and criteria to select the most appropriate network that answers user satisfaction and network efficiency. For example, a handover decision model designed from the user point of view is presented in (Calvagna et al., 2004). The authors propose two handoff decision policies (fixing a threshold value) between GPRS and WIFI networks. One of these policies aims to satisfy the user who is willing to pay for having its connections as granted as possible, while the other one tries to satisfy the user from connection cost point of view but will disappoint his expectation of QoS. In (Ormond et al., 2006), the authors give special focus to user satisfaction by using a utility function for non-real time applications such as FTP (file transfer). The network decision algorithm is based on the difference between the monetary value of data transferred and the real price charged with time completion prediction. The designed utility function uses decision metrics such as user's risk attitude (finding a compromise between paying less and accepting delays).

Fifth, handoff decision strategies based on cost functions focuses on evaluating each one of the networks that are willing to support user services. Handoff decision algorithms, in this case, can be expressed as a sum of weighted functions of specific parameters. In (Wang et al., 1999), a policy-enabled handoff decision algorithm is proposed along with a cost function that considers several context parameters. However, this cost function is very simple and cannot handle more sophisticated scenarios. In (Qiang et al., 2005), an adaptive multi-criteria handoff decision algorithm for radio heterogeneous networks was introduced. In (Zhang et al., 2003), a method that considers both RSS and bandwidth as two important parameters for the cost function was developed, although this investigation only considers a single RSS threshold which could cause a ping-pong effect.

4.3 Proposed context aware mechanism

As stated earlier, context awareness is an important task to be addressed in order to provide context information while triggering handoffs or selecting new network destinations. In fact, without prior knowledge, a mobile terminal must scan channels of different frequencies to discover existing nearby networks. As mentioned in (Wei et al., 2006), scanning 13 channels in 802.11b WLAN requires in excess of 400 ms. Moreover, to the best of our knowledge, all of the previous work dealing with context-awareness assume that context information can be obtained and exchanged through heterogeneous networks. Practically, operators and private networks are very reticent to the idea of sharing their context information. To cope with this important task, we introduce a context-aware approach that ensures information sharing and respects operator's privacy. In the rest of this section, we present the architecture, the logical modules and the context-aware procedure relevant to the proposed context-aware mechanism.

4.3.1 Architecture

Fig. 4.1 depicts the proposed architecture where two networks are connected to an IP backbone. Each network possesses a context-aware server (CAS) which manages local context information. Every CAS is identified by a "CAS_identifier" which is broadcasted through a periodic router beacons. For simplicity we consider that *network 1* and *network 2* are respectively connected to an IP backbone via CAS1 and CAS2. We also introduce an interworking cooperation server (ICS) that ensures context information sharing between heterogeneous technologies. The ICS unit should be owned by an independent authority or operator. We also assume that both *network 1* and *networks 2* have a registration entry with the ICS. This means that their respective context aware servers (CAS1 and CAS2) can periodically and securely exchange context information with the ICS. Notice that the architecture bellow can be easily extended to more than two networks since the ICS manipulates only signaling traffic.

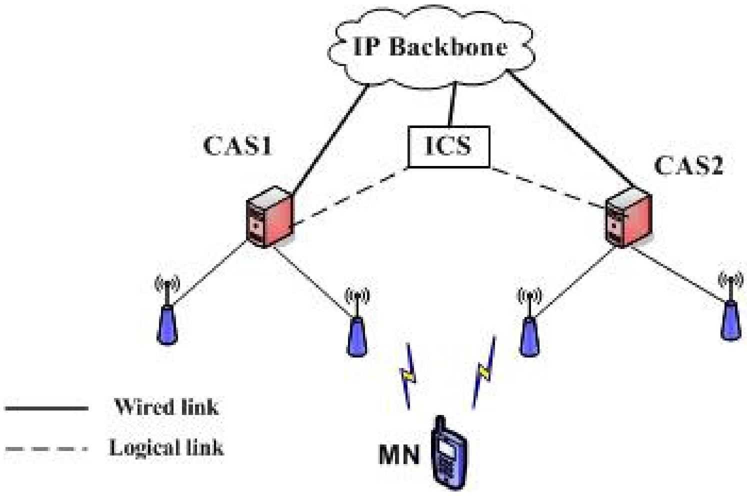


Figure 4.1 Context-aware architecture

4.3.2 Context-aware logical modules

The logical modules relevant to the MN, CAS and ICS are illustrated in Fig. 4.2.

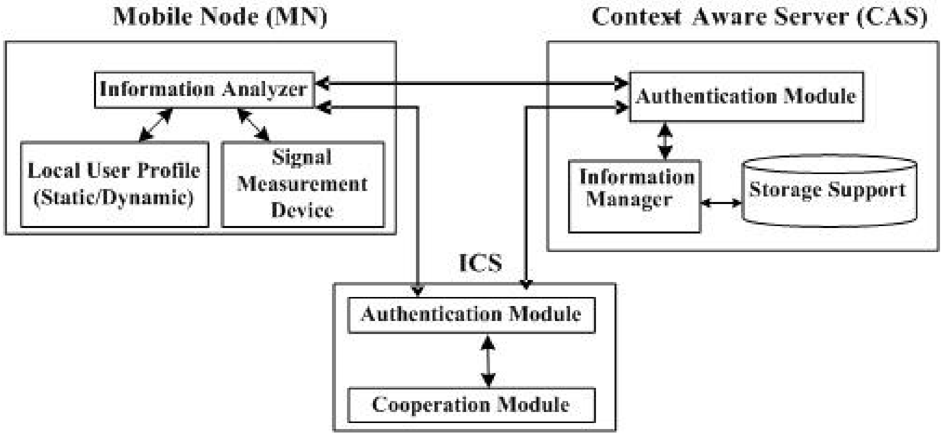


Figure 4.2 Context-aware logical modules

More specifically, each module operates as follows:

- *The Information Analyzer (IA)*

The main role of this module consists in managing local context information, performing handoff triggering and selecting new network destinations.

- *The Signal Measurement Device (SMD)*

The SMD module performs RSS measurements and continuously updates the local user profile (LUP).

- *Local User Profile (LUP)*

The LUP unit operates as a local database that stores both static and dynamic context information relevant to the MN. The static information may concern wireless card type, public encryption keys, etc., and dynamic information may concern MN's velocity, mobility patterns, RSS measurements, moving history, etc.

- *The Authentication Module (AM)*

The AM refers to the entity that communicates with external components and authenticates mobile users.

- *The Information Manager (IM)*

The information manager (IM) manages local context information (inside a subnet or network) and sends periodic context information to update CAS profile at the ICS. In this case, context information may refer to residual bandwidth, traffic status, connection blocking rate, etc.

- *Storage Support (SS)*

The main role of this entity is to store local network context information. However, it can also manage basic operations such as deleting obsolete information and providing novel data structures for new ones.

- *Cooperation Module (CM)*

This unit manages MN's requests and cooperates with distributed CASs to get accurate context information. Additionally, the CM allows QoS mapping between various mobile technologies. Mapping is needed to translate the QoS guarantees and specifications provided for a session across heterogeneous systems. The QoS mapping performed by this unit is for instance the requirements relevant to resource reservation subjected to the pre-established service level agreements (SLAs) between networks.

4.3.3 Context-aware procedure

When an MN needs context information from its nearby networks, it sends a *context_req* message to the ICS. This message contains a list of context information to be provided as well as the identifiers of the MN's neighbor CASs. The ICS authenticates the MN and sends a *context_get_infos* to the entire CASs located in the MN's vicinity. Each CAS replies to the ICS with a *context_infos_rep* that contains the requested context information. Finally, the ICS sends a *context_rep* to the MN. Fig. 4.3 shows the message flow in the presence of *n* CAS servers.

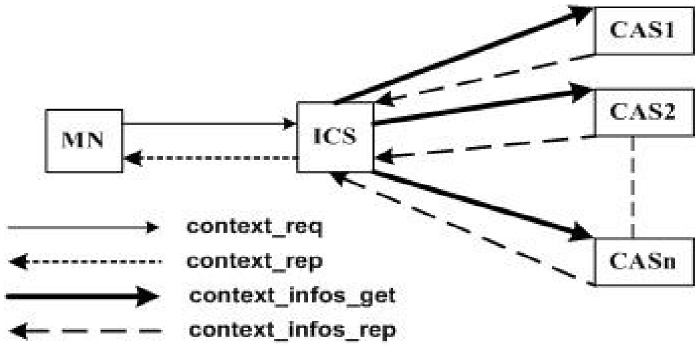


Figure 4.3 Context-aware flow messages

In this way the MN reduces signaling traffic in the wireless link since it avoids requesting individually all of its neighbors for context information. Moreover, delays relevant to MN's authentication, with each CASs, are avoided because the requested context information is obtained through the ICS which is assumed to have a secure entries with all the CASs.

4.4 Proposed Handoff Decision Strategy (HDS)

According to the precedent literature review, it was noticed that handoff initiation and network selection are still a challenging issue since RSS remains the most popular criterion used for these two tasks. In this section, we propose a new handoff decision strategy that considers more efficient context information while dealing with handoff initiation and network selection. More specifically, we first give an overview of the proposed strategy, and then we provide a detailed description of its main components.

4.4.1 Handoff decision strategy overview

The proposed handoff decision strategy, referred to as HDS, is illustrated in Fig. 4.4 HDS combines context awareness, fuzzy logic and score preference estimation to provide an adaptive approach that deals efficiently with handoff triggering and network selection.

Prior to handoff, an MN periodically obtains context information from its current home network. It can also obtain, when necessary, context information from its neighbor networks through the ICS. This information may concern link quality, signal strength, bandwidth, network's capabilities, subnet load, etc. This information is provided by the context aware scheme presented in section III. The received context information is fed into a handoff triggering scheme that uses fuzzy logic to decide whether the MN has to initiate forced or voluntary handoffs or simply remain in its current attachment point. We remember that forced handoffs are initiated in the case of MN's QoS deterioration while voluntary handoffs are triggered when the MN looks for new QoS conditions which are not available in its current home network. However, when an MN initiates a forced handoff, it verifies first whether it can perform a Layer 2 handoff toward an AP (access point) that satisfies its QoS requirements; otherwise, it starts the network discovery phase.

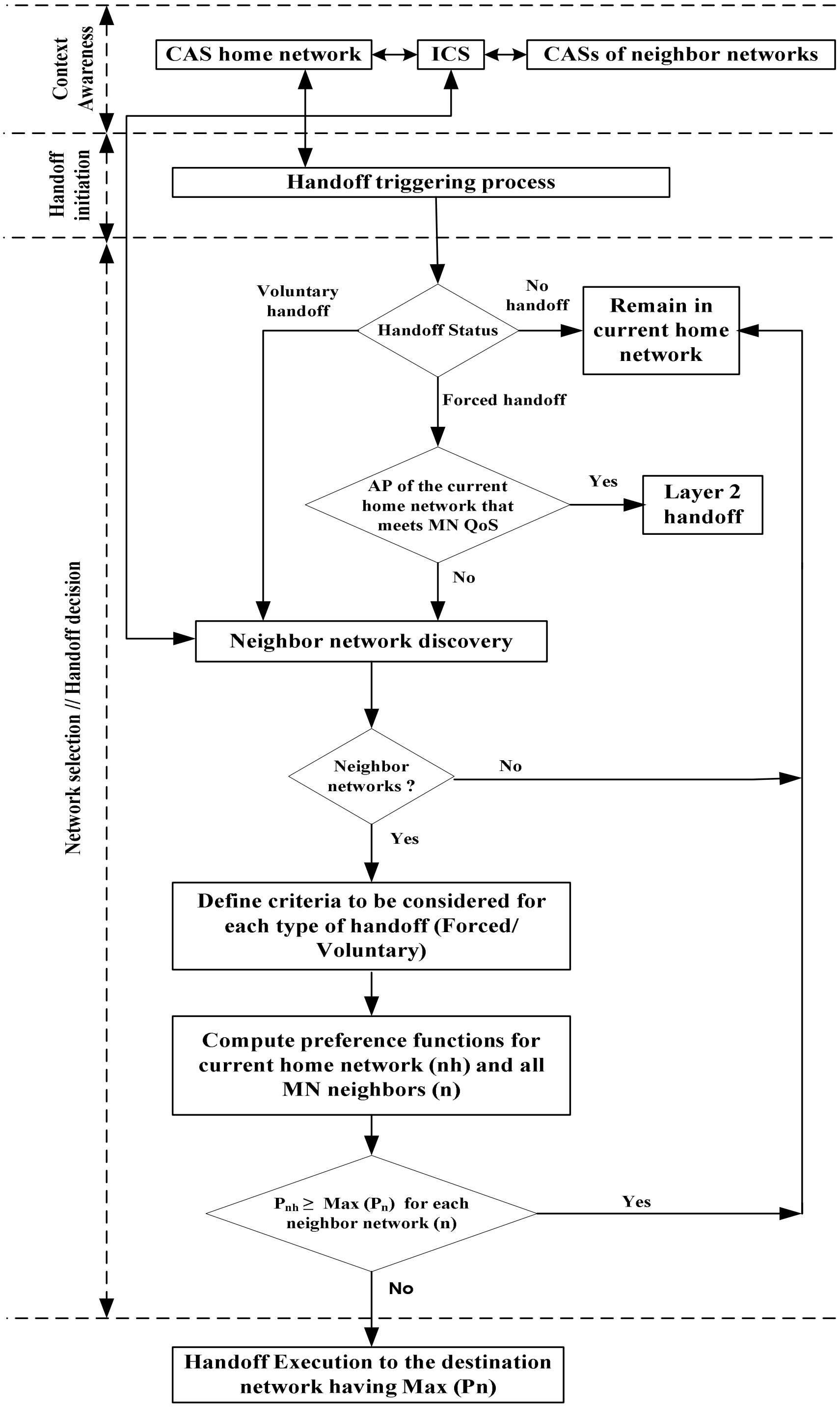

Context Awareness
CAS home network
ICS
CASs of neighbor networks
Handoff initiation
Handoff triggering process
Network selection // Handoff decision
Voluntary handoff
Handoff Status
No handoff
Remain in current home network
Forced handoff
AP of the current home network that meets MN QoS
Yes
Layer 2 handoff
No
Neighbor network discovery
Neighbor networks ?
No
Yes
Define criteria to be considered for each type of handoff (Forced/ Voluntary)
Compute preference functions for current home network (nh) and all MN neighbors (n)
$P_{nh} \geq$ Max (P_n) for each neighbor network (n)
Yes
No
Handoff Execution to the destination network having Max (Pn)

Figure 4.4 Flow chart of the proposed handoff decision strategy (HDS)

Neighbor network discovery is also initiated when the MN triggers a voluntary handoff. At this stage, the MN aims to find out a list of eventual candidate networks that will satisfy its QoS needs. However, to avoid all air-interfaces always on approach for system discovery, we propose an adaptive scheme. An MN requests neighbor networks information from its serving network to the ICS. Trough information reported periodically to the ICS, it maintains a global view of the connection state of roaming MNs and access network conditions in its coverage area. The ICS replies by sending information about neighbor networks to the MN through its point of attachment.

Once the MN constitutes a list of candidate networks that can eventually guarantee its QoS requirements, it defines a complete set of criteria to be considered depending on the initiated handover (forced vs voluntary). In addition to the type of criteria, this set includes thresholds (i.e. minimum QoS requirements) and weights relevant to the considered context parameters. Then, the MN estimates a preference score function for each one of the candidate networks to decide where to handoff. Finally, handoff execution takes place to effectively re-establish MN's connections and complete the inter-system roaming process.

4.4.2 HDS main modules

The main components of the proposed HDS are: context awareness, handoff initiation and network selection. We remember that the context awareness module has been already presented in section III, so its description will be skipped here. Thus, in the rest of this subsection, we focus on handoff triggering and network selection.

4.4.2.1 Handoff triggering

Handoff triggering is a crucial issue since the MN must decide which type of handoff to initiate and which context parameters to consider for that purpose. As stated before, the received signal strength is not enough to trigger efficient handoffs, i.e., on right time and under appropriate parameters. Thus, we advocate that handoff initiation should take into account various context criteria. However, it is difficult to define handover triggering conditions since context parameters can be expressed both in crisp and linguistic values. That is why we propose a fuzzy logic based solution that initiates handovers with different context parameter types (crisp, linguistic, etc.). In Fact, fuzzy logic (Zadeh, 1972) is a powerful concept that uses imprecise and uncertain data to produce precise values and actions. This is advantageous in the target networks because a fuzzy logic system is flexible and can be used to model nonlinear functions with arbitrary complexity.

As it is shown in Fig. 4.5 the first step of the proposed handoff initiation scheme consists in feeding the received context parameters into a fuzzifier. The main role of the fuzzifier is to transform real-time measurements into fuzzy sets, which contain elements with different membership degrees. For example, if the RSS signal is considered in a crisp set, it can be either weak or strong. However, in a fuzzy set, the RSS signal can be considered as quite weak, medium or strong. Membership values are generated by mapping the values obtained for particular parameters onto a membership function like the ones illustrated in Fig. 4.6. In general, these functions consist of a curve or line that defines how each datum or value is mapped onto a membership value. For instance, in Fig. 4.6 (a) S_1 is assigned the value 0.6 in the *Almost weak* set, 0.3 in the *weak* set and 0 in the *Medium* and *Strong* sets.

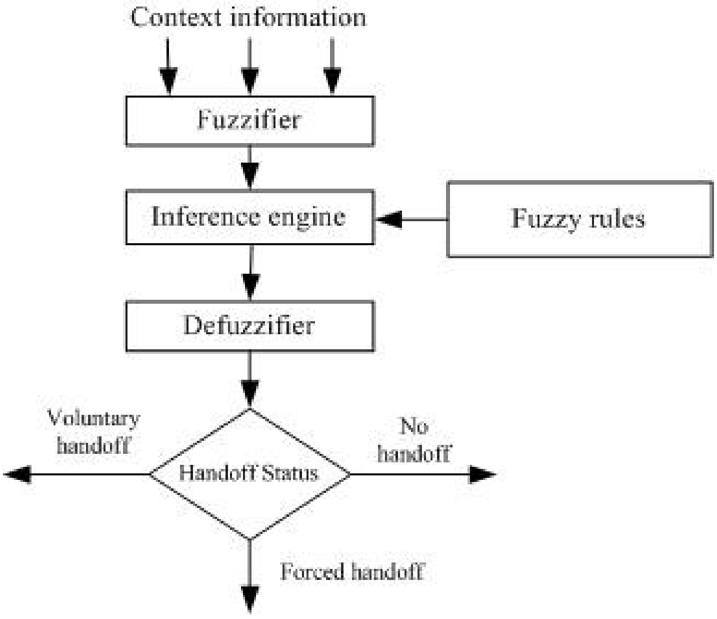
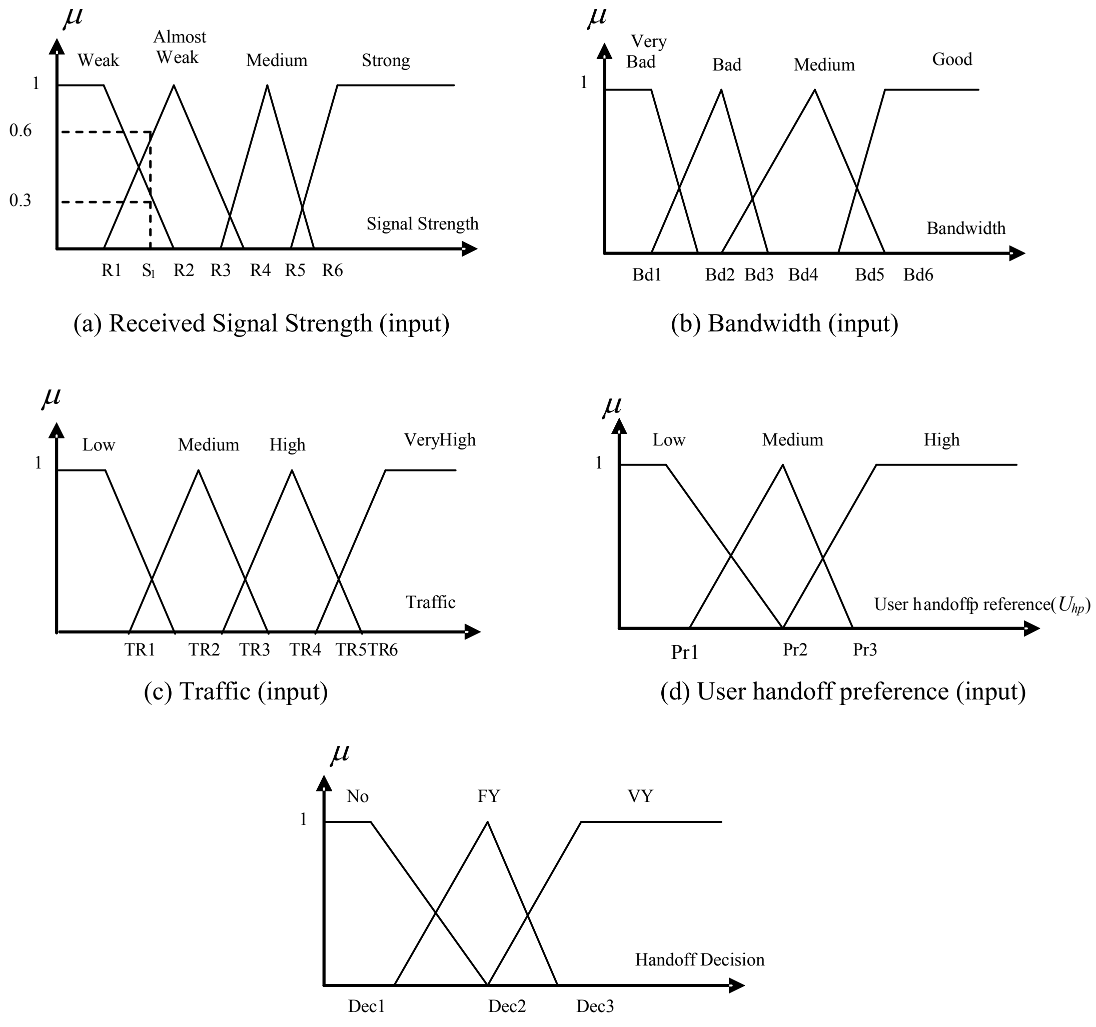


Figure 4.5 Handoff triggering process

The second step of handoff initiation involves feeding the fuzzy sets into an inference engine, where a set of fuzzy IF-THEN rules is applied to obtain fuzzy decision sets. Fuzzy rules can be defined as a set of possible scenarios which determine whether a handover is necessary or not. The proposed initiation mechanism basically considers three decision sets: Forced Yes (FY), Voluntary Yes (VY) and No handoff (N). An example of IF-THEN fuzzy decision rules appears in Table 4.1. The output fuzzy decision sets are aggregated into a single fuzzy set and sent to the defuzzifier to be converted into a precise quantity during the last step of the handover initiation using the centroid method (Chi-Hsing et al., 1999).



(a) Received Signal Strength (input)

(b) Bandwidth (input)

(c) Traffic (input)

(d) User handoff preference (input)

(e) Handoff decision (output)

Figure 4.6 Example of membership functions

Table 4.1 Example of Fuzzy Rules

	Fuzzy Rules
Rule 1	IF (RSS is weak) or (Traffic is very high) or (Bandwidth is very bad) THEN Handoff is FY
Rule 2	IF (Bandwidth is bad) and (RSS is almost weak) THEN Handoff is FY
Rule 3	IF (Bandwidth is bad) and (Traffic is high) THEN Handoff is FY
Rule 4	IF (RSS is almost weak) and (Traffic is high) THEN Handoff is FY
Rule 5	IF (Bandwidth is medium) and (User handoff preference is high) THEN Handoff is VY
Rule 6	IF (Traffic is high) and (User handoff preference is high) THEN Handoff is VY
Rule 7	IF (RSS is medium) and (User handoff preference is high) THEN Handoff is VY
Rule 8	IF (RSS is almost weak) or (Bandwidth is bad) and (Traffic is high) and (Handoff preference is not Low) THEN Handoff is VY
Rule 9	IF (RSS is not almost weak) and (Bandwidth is not bad) and (Traffic is not high) and (Handoff preference is not high) THEN Handoff is N
Rule 10	IF (Handoff preference is low) THEN Handoff is N
Rule 11	IF (RSS is not weak) and (Bandwidth is not very bad) and (Traffic is not very high) and (Handoff preference is not high) THEN Handoff is N

4.4.2.2 Network selection

Network selection or handoff decision making refers to the process of choosing the most suitable network destination that satisfies MN's requirements in terms of QoS, monetary cost, security, battery consumption, user preferences, etc. Practically, this process passes though: neighbor network discovery, context preparation and score function calculation.

a) Neighbor network discovery

This phase consists in finding out all of the MN's neighbor networks which are willing to support its ongoing services. In fact, this task can be completed through the periodic beacons which include identifiers pertaining to MN's neighbor CASs. Otherwise, the MN sends a *neighbor_infos_req* message to the ICS which maintains a global view of the entire integrated mobile systems. The ICS replies with a *neighbor_infos_rep* message which contains a list of MN's neighbor networks.

b) Context preparation

Here, the MN defines its context criteria depending on the type of handoffs to be initiated (forced or voluntary). Moreover, it specifies context criteria thresholds, i.e., minimum QoS requirements as well as their corresponding weights.

c) Score function calculation

The next generation of mobile networks (4G) aims to guarantee ongoing communications through heterogeneous mobile technologies. However, the selection of network destination that provides subscribers with better services remains a challenging issue and depends on several parameters such as : bandwidth, power consumption, user preferences, monetary cost, type of handoffs (forced vs voluntary), network stability, etc. The design of a handoff decision function that takes into account these parameters is crucial and needs a consensus between user requirements and network capabilities.

In this section, we propose a new handoff score function that allows a best network selection based on wide range of context parameters including network stability. More specifically, it models the relationship between user services and network capabilities for both forced and voluntary handoffs. This means that the proposed handoff score function estimates network destination preferences depending on the type of the triggered handoff (forced or voluntary). In fact, forced handoffs need quicker network selection since mobile users have to complete immediately their handover process. Therefore, the selected network destination is performed under minimum context parameters. In case of voluntary handoffs, network selection is performed with a large number of context variables. In the following we present the proposed score function as well as its relevant calculation procedure.

● Preference function definition

Let C^F and C^V denotes respectively the sets of criteria used to select networks in case of forced and voluntary handoffs.

In the rest of this section, mobile user and mobile node (MN) will be used interchangeably.

For a given mobile user *u*, we define a best network destination as:

$$Network\ n^* = \underset{n \in N}{Max}\{P_u^n\} \qquad (1)$$

Where:

P_u^n refers to the estimated preference for network *n* to run on user services,

N denotes the set of neighbor networks.

P_u^n is defined by:

$$P_u^n = \begin{cases} P_{u,forced}^n & \text{if mobile user is subjected to a forced handoff} \\ P_{u,voluntary}^n & \text{if a mobile user is subjected to a voluntary handoff} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

$P_{u,forced}^n$ and $P_{u,voluntary}^n$ are respectively given by :

$$P_{u,forced}^n = (\sum_{i=1}^{m}\sum_{j=1}^{q}\omega_{s_i,c_j} \cdot P_{u,c_j}^{n,s_i}) \cdot R_{c,f}^n \quad , \quad c_j \in C^F \qquad (3)$$

$$P_{u,voluntary}^n = (\sum_{i=1}^{l}\sum_{j=1}^{r}\omega_{s_i,c_j} \cdot (2 \cdot P_{u,c_j}^{n,s_i} - P_{u,c_j,precedent}^{n,s_i})) \cdot R_{c,v}^n \ , \quad c_j \in C^V \qquad (4)$$

Where :

ω_{s_i,c_j} denotes the weight to meet service s_i under criteria c_j,

$\omega_{s_i c_j} \in [0,1]$, $\sum_{i,j}\omega_{s_i,c_j} = 1$,

P_{u,c_j}^{n,s_i} refers to the estimated preference to meet a user service s_i on network *n* under criteria c_j,

$P^{n,s_i}_{u,c_j,precedent}$: refers to the precedent estimated preference to meet a user service s_i on network *n* under criteria c_j,

$R^n_{c,f}$ and $R^n_{c,v}$ denote respectively factors relevant to forced and voluntary handoffs, which are used to eliminate networks that do not meet user's requirements. They are defined by :

$$R^n_{c,f} = \begin{cases} 1 & \text{if } \prod_{i,j} P^{n,s_i}_{u,c_j} \neq 0 \qquad (i=1,.., m) \text{ and } (j=1,..,q) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$R^n_{c,v} = \begin{cases} 1 & \text{if } \prod_{i,j} P^{n,s_i}_{u,c_j} \neq 0 \qquad (i=1,.., l) \text{ and } (j=1,..,r) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Finally, to assess the stability of candidate networks and avoid the ping-pong effect, the term $: 2 \cdot P^{n,s_i}_{u,c_j} - P^{n,s_i}_{u,c_j,precedent}$ (Equation (4)) is used to penalize instable networks and support the ones which improve QoS parameters. Indeed, the term $2 \cdot P^{n,s_i}_{u,c_j} - P^{n,s_i}_{u,c_j,precedent}$ is equivalent to $P^{n,s_i}_{u,c_j} + P^{n,s_i}_{u,c_j} - P^{n,s_i}_{u,c_j,precedent}$. Hence, if P^{n,s_i}_{u,c_j} is greater than or equals to $P^{n,s_i}_{u,c_j,precedent}$, the final score will be improved or remain stable (in case of equality), otherwise the final score will decrease.

• Preference function computation

The proposed preference function can be computed either at MN or at the ICS side. In fact, if we assume that mobile devices will become increasingly powerful, intelligent and sensitive to link layer changes we can adopt network assisted and mobile-controlled handoff strategy. This means that networks provide context information and the MN estimates their relevant preference functions to decide where to handoff. On the other hand, if MN capabilities are limited we will adopt mobile-assisted and network controlled strategy. In other words, the MN will provide it service's criteria in terms of QoS parameters, preference requirements (weights) and thresholds (i.e., minimum QoS), then the ICS computes the preference function pertaining to each candidate networks. We advocate that this last approach will allow the MN to save resources in terms of computing time and energy consumption. Moreover, the privacy of network's context information is respected since the MN will receive only results of score preferences relevant to each neighbor network rather than manipulating their context information.

4.4.2.3 Handover execution

The main concern of this module is to ensure service continuity while roaming through heterogeneous mobile systems. This task can be completed by Mobile IP (Johnson et al., 2004) based solutions such as HMIP (Soliman et al., 2005), FMIP (Koodli, 2005), FHMIP (Jung et al., 2005), etc. It can also be completed at the transport layer by SCTP (Stewart, 2007) based mobility schemes that use multihoming and dynamic address reconfiguration features.

4.5 Simulations and Results

In this section, we study the effectiveness of the proposed handoff decision strategy (HDS). To complete this task, we choose RSS based handoff decision strategy as a comparison benchmark since the RSS parameter is widely used in many previous work and systems (Lassoued et all., 2008). More specifically, we first present the used simulation model, and then we discuss the obtained results.

4.5.1 Experimental Model

Fig. 4.7 depicts the simulation model used for performance analysis. Each BS_i refers to *network i* (i.e. *operator i*) which is supposed to use the same physical layer technology. We assume that BS_1 is enhanced with ICS features while the rest of BS_i (i = 2 to n) are endowed with CAS functionalities. In each experiment, the MN is assigned to *network 1* (i.e., BS_1) and moves in a constant speed from a start position (S) until the end position (E) located in the overlapping area as it is shown in Fig. 4.7.

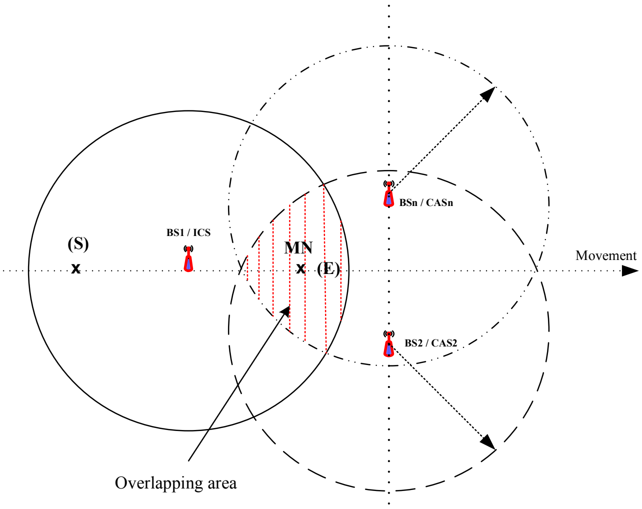


Figure 4.7 Simulation model

All simulations are completed according to the process illustrated in Fig. 4.8. More specifically, this process starts by an initiation setup which consists of generating n overlapping networks as illustrated in Fig 4.7. When the MN reaches the overlapping area, we produce a random deterioration of the QoS parameters (i.e., RSS, bandwidth and traffic status) relevant to the MN's home network. Then, a Fuzzy Logic based triggering procedure is launched to decide whether to initiate a forced or a voluntary handoff. Depending on the type of the handoff to be triggered, each BS_i sends a list of context parameters to the ICS. Then, the ICS estimates a preconfigured preference function (depicted in section IV) for each *Network$_i$*. After, a list of Candidate Network destination is sent to the MN. Finally, the MN selects the destination having the maximum score result. The context parameters and their corresponding weights, considered in our experiments, are shown respectively in Table 4.2 and 4.3. The overall execution simulation process is outlined in Fig 4.8 and it is implemented in C++ and uses the Matlab Fuzzy tool.

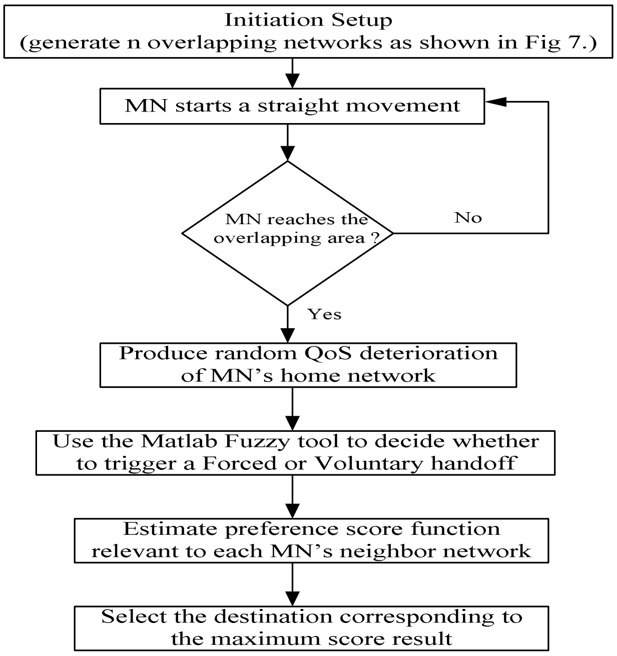

Initiation Setup
(generate n overlapping networks as shown in Fig 7.)
MN starts a straight movement
MN reaches the overlapping area ?
No
Yes
Produce random QoS deterioration of MN's home network
Use the Matlab Fuzzy tool to decide whether to trigger a Forced or Voluntary handoff
Estimate preference score function relevant to each MN's neighbor network
Select the destination corresponding to the maximum score result

Figure 4.8 Simulation execution process

The normalized preferences used for score calculations are defined by:

$P^{n,s_i}_{u,RSS} = \dfrac{RSS_n}{10}$, $RSS_n \in [0,10]$;

$P^{n,s_i}_{u,bandwidth} = 1 - e^{-B_n}$, B_n refers to the residual bandwidth on network n, $B_n \geq 0$;

$P^{n,s_i}_{u,price} = e^{-C_n}$, C_n refers to service cost per min, $C_n \geq 0$;

$P^{n,s_i}_{u,battery} = e^{-P_n}$, P_n equals to power consumption per hour, $P_n \geq 0$;

$P^{n,s_i}_{u,sojourn} = 1 - e^{-S_n}$, S_n refers to sojourn time per MN visit, $S_n \geq 0$.

$P^{n,s_i}_{u,traffic} = e^{-T_n}$, T_n indicates traffic status on network n, $T_n \geq 0$.

The above context parameters can also be expressed in a quotient form i.e., $\frac{\left(X_n^{Max} - X_n\right)}{X_n^{Max}}$ (X_n refers to a context criterion *n*), however we will use the exponential form since it is easy to handle and avoids singularities while generating random values.

Table 4.2 Context criteria relevant to HDS and RSS-based handoff strategies

Type of handoff strategies				Context parameters
RSS-based	Hanoff triggering		:	*RSS* of home network
	Network selection		:	*RSS* of neighbor networks
HDS	Handoff triggering		:	*RSS*, *Bandwidth* and *Traffic status* of MN's home network
	Network selection	Forced	:	*RSS*, *Bandwidth* and *Traffic status* of MN's neighbor networks
		Voluntary	:	*RSS*, *Bandwidth*, *Traffic status*, *Monetary cost*, *Power consumption*, *Sojourn time* of neighbor networks

Table 4.3 Example of service weights

Criterion	RSS	Traffic	Bandwidth	Price	Sojourn time	Battery
Normalized voice weights	0.225	0.125	0.175	0.2	0.15	0.125
Normalized download weights	0.162	0.109	0.216	0. 216	0.162	0.135

4.5.2 Results

This section presents and discusses results relevant to the use of the proposed HDS and the RSS-based handoff strategies while performing handovers through heterogeneous networks. More specifically, we investigate the impact of handoff initiation type (forced *vs* voluntary) on the quality of selected network destination.

Fig. 4.9 illustrates the estimated preference score relevant to the selected network destination as a function of the number of MN's neighbor networks (*BS*$_i$). The first thing to be noted is that, the entire networks selected during voluntary handoffs are associated with high preference scores compared to the RSS-based handoffs. Such a situation is quite normal since during a voluntary handoff, the MN destination corresponds to the one that maximizes score preference under several context parameters. Therefore, the chosen destination meets all of the MN requirements, such as high bandwidth, maximal sojourn time, minimal financial costs, etc. On the other hand, the RSS-based handoffs select only networks that meet high RSS values. However, this type of selected network destination may have, for instance, poor bandwidth, low sojourn time, high monetary cost, etc. That is why the chosen networks under RSS comparisons show less score preferences compared to the ones selected in case of voluntary handoffs. Nevertheless, 9.4% of the cases under investigation show that RSS-based handoffs can also have good network scores. This situation is particularly evident when the number of base stations is low.

In fact, let N_{BS} be the number of MN's nearby networks and N_v the number of context parameters considered for a voluntary handoff. Let $P_{Succ}^{RSS,V}$ be the success probability that an MN, subjected to a RSS-based handover, chooses a same network destination as it performs a voluntary handoff. In other words, $P_{Succ}^{RSS,V}$ refers to the probability that the preference score relevant to an RSS handoff will be equal to the one estimated for a voluntary handover. This probability is expected to reach high values when N_{BS} is low (limited choice for network destination). Subsequently, $P_{Succ}^{RSS,V}$ will decrease as the number of nearby networks increases. Moreover, when the number of context criteria associated with a voluntary handoff (N_v) is low, $P_{Succ}^{RSS,V}$ is expected to increase. Particularly, when N_v = 1, the RSS criterion will be the only context parameter used for voluntary handoffs. In this case, voluntary and RSS-based handoffs will select the same network destination (i.e., $P_{Succ}^{RSS,V}=1$).

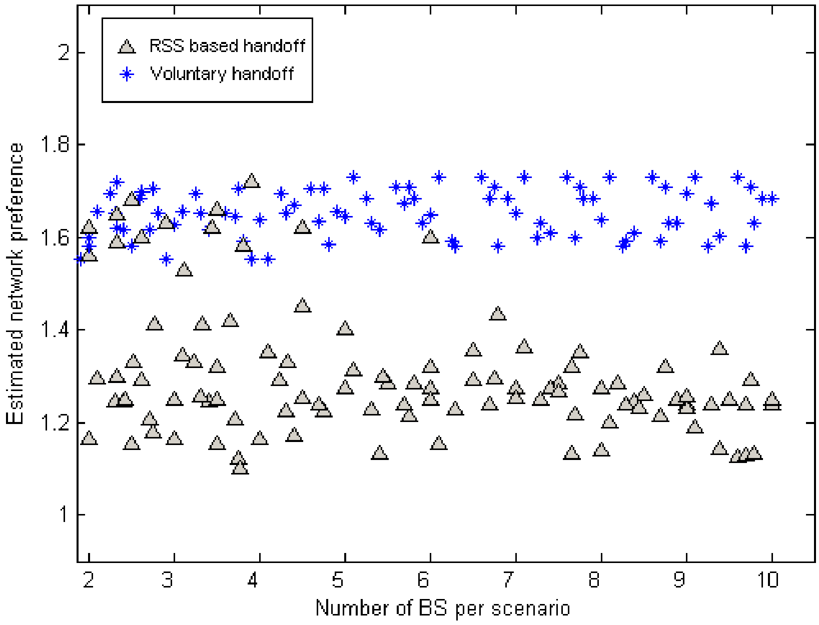


Figure 4.9 Comparison of RSS vs. voluntary based handoffs

We have analyzed the $P_{Succ}^{RSS,V}$ behavior during several experiments, and we have noticed that this probability decreases rapidly when N_{BS} and N_{v} increase. Hence, we approximate the RSS success probability of meeting high network scores by: $P_{Succ}^{RSS,V} = \frac{1}{e^{\alpha(N_{BS}-1)}}$.

Where α refers to a decreasing factor which is introduced to illustrate the impact of the number of context criteria on the RSS success probability. The decreasing factor α can be expressed by: $\alpha=(N_{v}-1)/N_{v}$. Accordingly, when voluntary handoffs consider only the RSS context criterion (α=0), $P_{Succ}^{RSS,V}$ equals to 1; otherwise, it decreases according to the number of BSs (N_{BS}) and voluntary context criteria (N_{v}), as shown in Fig. 4.10.

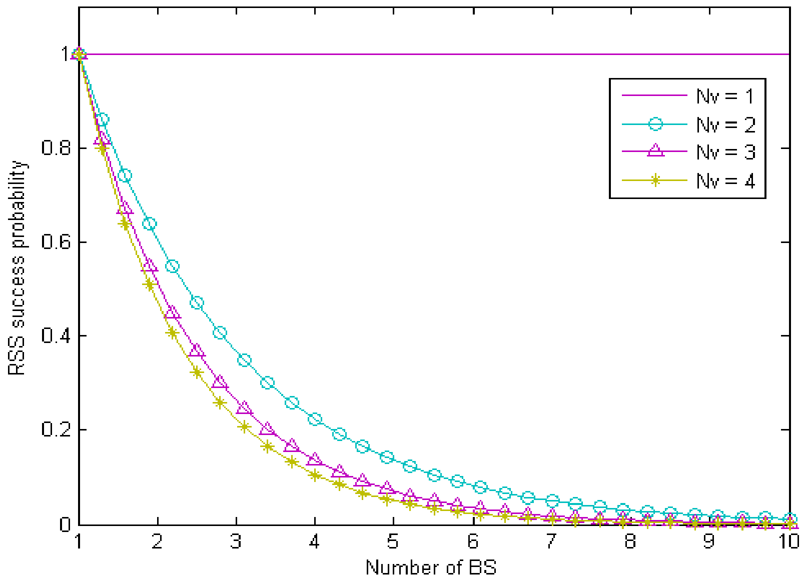


Figure 4.10   RSS probabilities of reaching high network scores

Fig. 4.11 illustrates score preferences relevant to the selected network destination when the MN is subjected to both RSS-based and forced handoffs. We notice that networks chosen when using forced handoffs show generally high preference scores compared to RSS-based handoffs. This is because when an MN performs a forced handoff, the proposed HDS allows it to consider at least 3 context parameters (e.g. RSS, bandwidth and traffic status) for network selection. Thus, the selected network destination satisfies MN's requirements in terms of RSS, bandwidth and traffic conditions. This is completely different from RSS-based strategy which tries to select a network destination that presents good RSS and remains unaware about the rest of MN's requirements.

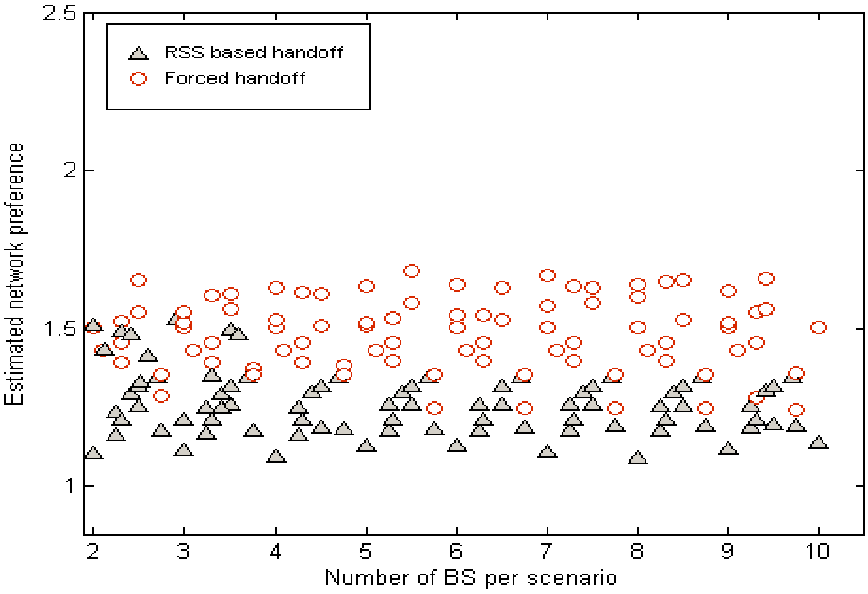


Figure 4.11 Comparison of RSS vs. force-based handoffs

Fig. 4.12 shows results pertaining to the estimated preference score for both forced and voluntary handoffs. Notice again that, voluntary handoffs allow the MN to roam to high score networks. This means that the selected network destination, using a voluntary handoff, supports all of the user services and ensures better network parameters compared to force-based handoffs.

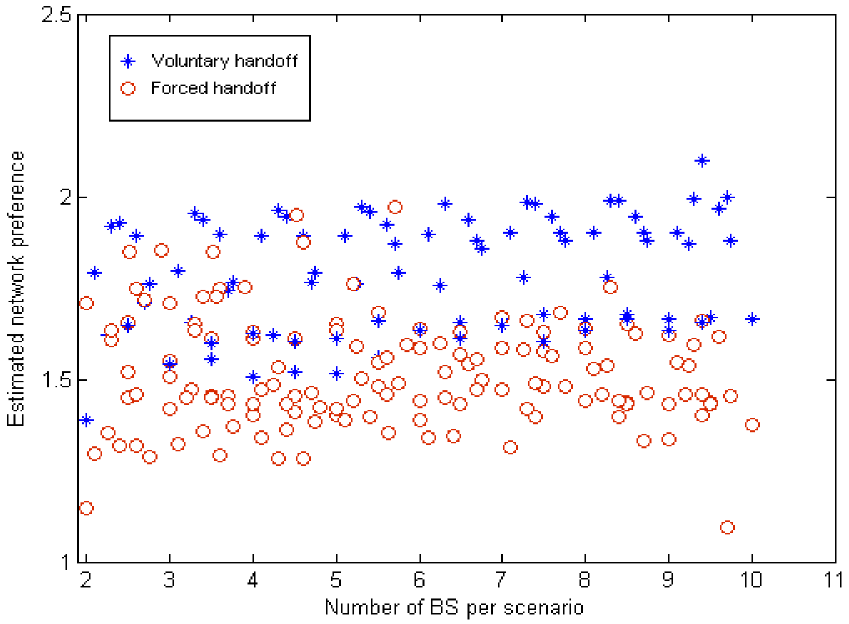


Figure 4.12 Comparison of voluntary vs. force-based handoffs

Similarly to the RSS success probability, we define the forced success probability to meet voluntary preference scores by: $P_{Succ}^{forced,V} = \frac{1}{e^{(N_{BS}-1)(N_V-N_F)/N_V}}$,

where

N_F : refers to the number of forced handoff criteria,

N_V : refers to the number of voluntary handoff criteria,

and N_{BS} : indicates the number of MN's nearby networks.

As we can see in Fig. 4.13, N_v is fixed at 6 and the number of context parameters relevant to the forced handoff (N_F) varies from 2 to 6. When N_F is low, the forced success probability is low as well. This situation is particularly observed when the number of BSs increases as it is shown in Fig. 4.13. However this probability becomes more and more important when N_F is quite close to N_V and the number of nearby BSs is low.

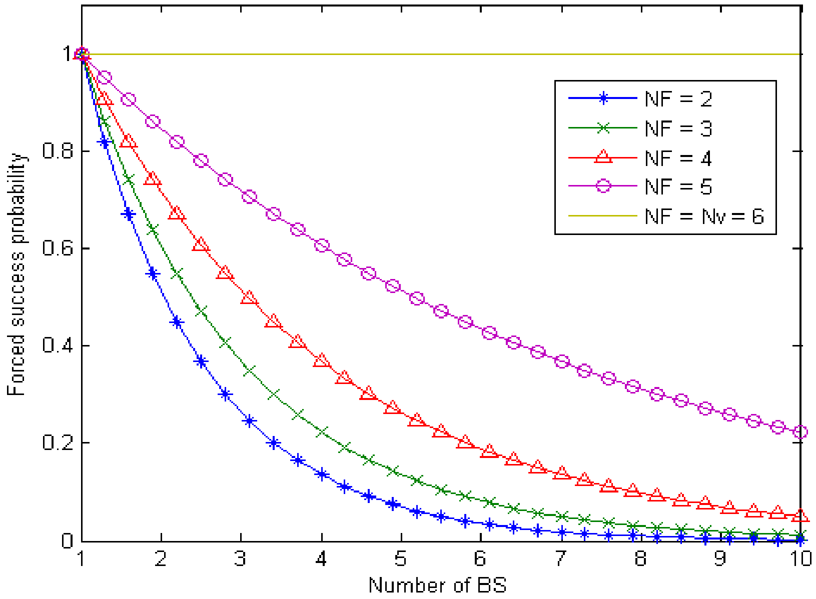


Figure 4.13 Forced success handoff probability to meet high network scores

In order to observe the difference between the compared handoff strategies, Fig. 4.14 illustrates the average network score as a function of the number of MN's neighbor networks (BSs). The average network score is calculated from the cumulative results obtained during several experiments. The first noticeable aspect resides in the fact that the average score is generally most important in the presence of a small number of nearby networks for both RSS and HDS based handoff strategies. However, as the number of nearby base stations increases, the average score pertaining to RSS-based and forced handoffs decreases. On the other hand, the average score associated with voluntary handoffs remains approximately the same. Accordingly, we can state that the proposed handoff decision strategy improves significantly the quality of the selected network destination with respect to user requirements and network capabilities.

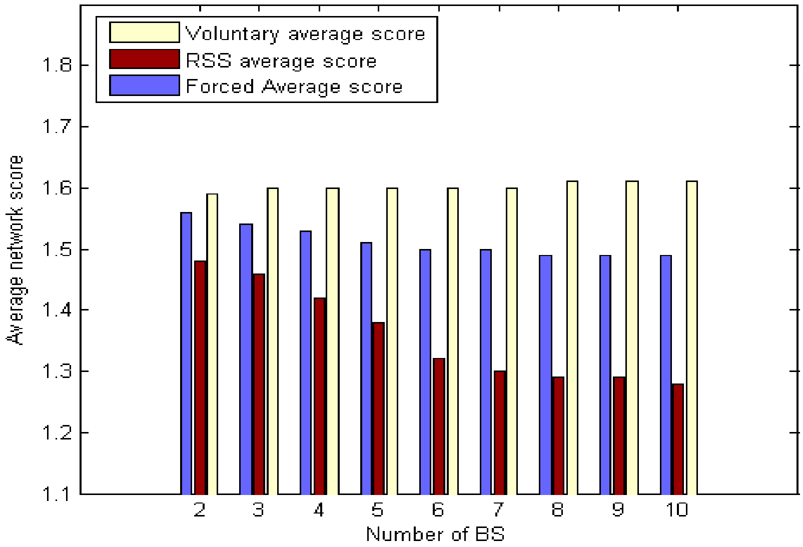


Figure 4.14 Average network score

To study the stability of the selected network destination when the MN is subjected to voluntary and RSS based handoffs. We define a stability factor as: $S_{fact} = e^{-\sum_{j} \frac{|c_j - c_{j,p}|}{c_j^{Max}}}$, where c_j refers to a context criterion *j*, $c_{j,p}$ indicates the previous value of criterion c_j and $c_{j,\max}$ corresponds to the maximum value of c_j. Fig. 4.15 shows the behavior of the stability factor for both voluntary and RSS based handoffs. We notice that the stability factor relevant to the voluntary handoff presents low fluctuations and remains approximately equal to one. This means that the selected networks, when the MN performs a voluntary handoff, do not present noticeable variations in the considered context criteria (difference between current and old values). These results is due to the fact that the score function, used in case of voluntary handoffs, takes into account network stability as it is shown in Eq (4). However, in the case of RSS-based handoffs, we notice more fluctuations in the stability factor. This means that the selected destination networks suffer from significant context variations which may lead to connection disruptions or performing handovers toward highly dynamic networks such as MANETs (Mobile Ad hoc NETworks).

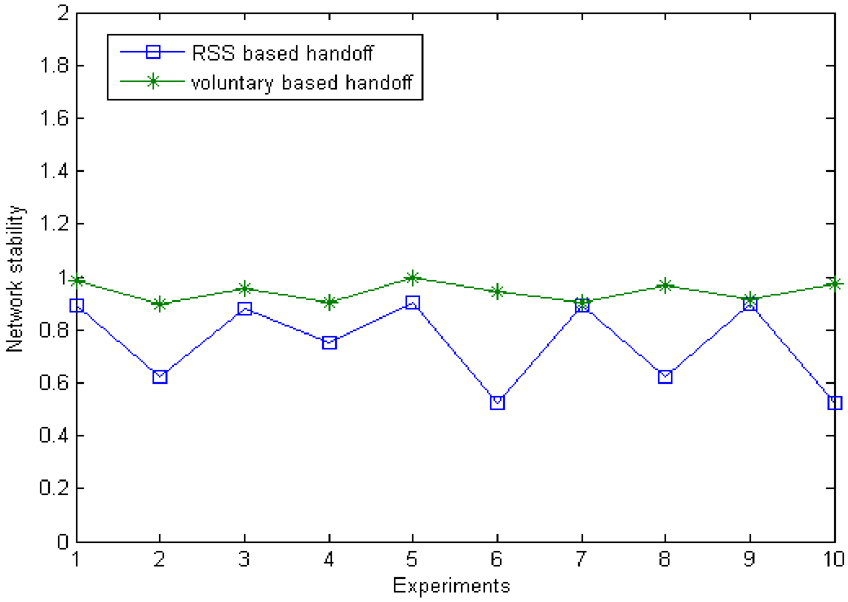


Figure 4.15 Estimated network stability for voluntary and RSS based handoffs

Now, let *n* be the number of MN's candidate networks (i.e., number of MN's neighbor CASs). Fig. 4.16 illustrates the average exchanged wireless messages as a function of number of handoffs. We notice that the use of the ICS (ICS-based), during preference scores computation, leads to significant reduction of the wireless link load (in terms of wireless messages) compared to the case where the MN calculates its score function through the context aware servers (CAS-based). Indeed, when the MN uses the ICS, the wireless link is solicited two times per handoff (refer to the computation score procedure introduced in (4.4.2.2 (c)). However, without the ICS, the MN has to exchange wireless messages with each one of its neighbor CASs. Thus, the wireless link load increases depending on the number of MN's candidate networks (i.e. *n*).

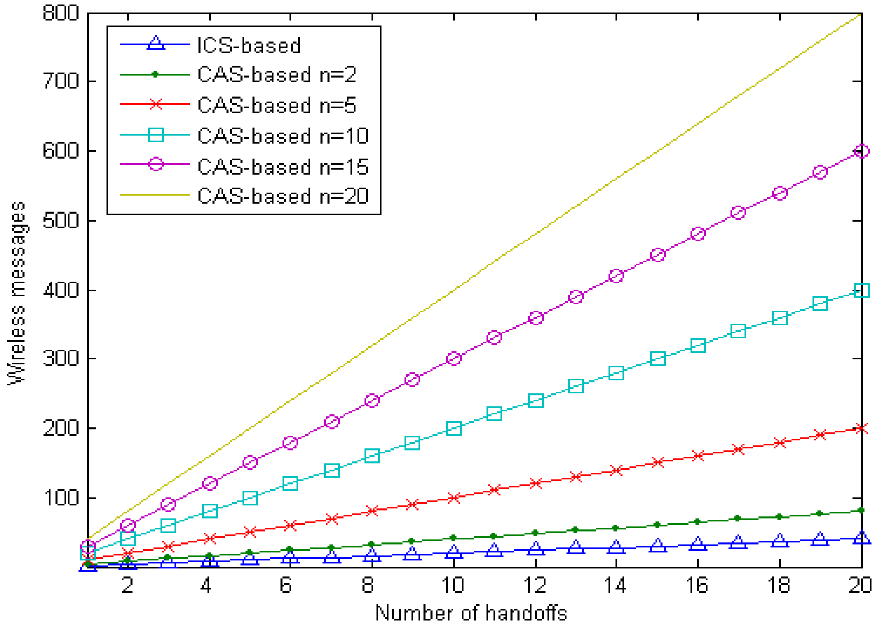


Figure 4.16 Exchanged wireless messages

4.6 Conclusion

This paper proposes a new handoff strategy that uses fuzzy logic and context-awareness to improve the handoff triggering processes and arrive at a more efficient choice of MN network destinations. Unlike traditional decision approaches, this novel solution considers a large number of context information such as price, RSS, bandwidth, sojourn time, power consumption, etc. Such criteria are managed by an efficient context-awareness mechanism. Furthermore, a preference function was defined to model the relationship between MN requirements and network capabilities. This function models two types of handoffs, defined as forced and voluntary. Then, a fuzzy-based handoff triggering approach was proposed to select which kind of handoff (forced, voluntary) is to be initiated. The results thus obtained show that, voluntary handoffs ensure better network destination as compared to forced and RSS-based handoffs. This investigation also shows that forced handoffs yield better results, as compared to RSS-based handoffs, since they generally guarantee high network scores. In future work, we intend to compare the obtained results with other score methods such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) and AHP (Analytical Hierarchical Process). Additionally, we propose to model the user preference toward a handoff which refers to the real need to perform or not an eventual handoff. We expect that the introduction of such parameter will avoid unnecessary handoffs and then participates to optimize both MN and network resources.

CHAPITRE 5

TOWARDS CROSS LAYER MOBILITY SUPPORT IN METROPOLITAN NETWORKS

Abdellatif Ezzouhairi, Alejandro Quintero, Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer Engineering, École Polytechnique de Montréal
C.P. 6079, succ. Centre-Ville, Montreal, Quebec, H3C 3A7, Canada
Phone: (514) 340-3240 ext. 4685. Fax: (514) 340-3240
Emails: {Abdellatif.Ezzouhairi, Alejandro.Quintero, Samuel.Pierre}@polymtl.ca

Abstract

The next generation mobile networks (4G) is expected to integrate a large number of heterogeneous wireless systems. Practically, metropolitan networks will play an important role in such integration since they include a great variety of mobile systems. Mobility management as well as the integration of existing/future wireless technologies remain an important task to be investigated. A number of interworking proposals are available in the literature, yet none can claim to be the ultimate and unique integrated solution. Moreover, these proposals fail to guarantee seamless mobility and service continuity. This paper proposes a Hybrid Interworking Architecture (HIA) that integrates mobile systems in a metropolitan area. Moreover, the proposed HIA is endowed with a cross layer mobility scheme that guarantees seamless roaming and service

continuity. Performance analyses show that our proposals exhibit net improvements of QoS guarantee compared to existing solutions.

Index Terms: Integration, architecture, hybrid, mobility management, metropolitan.

5.1 Introduction

In the last few years, metropolitan mobile networks have gained more attention since they reflect a concrete use of mobile systems within a city coverage. However, this category of networks becomes more and more heterogeneous and depends on various mobility approaches. Actually, a metropolitan network may include: WiFi for short distance wireless local area networks (WLANs), UMTS/cdma2000 for broadband on 3G cellular networks, Bluetooth for personal area coverage. An illustrated example of metropolitan networks is given in Fig. 5.1.

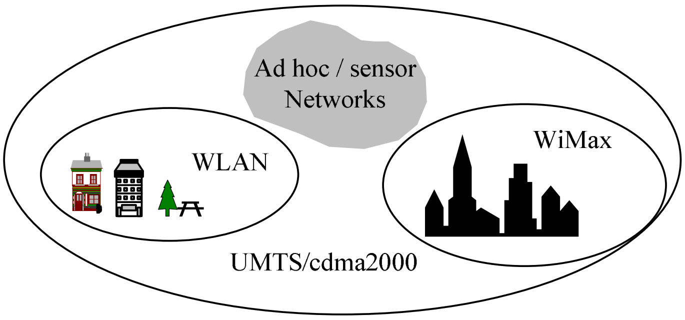


Figure 5.1 Example of a Metropolitan network

Practically, each technology is conceived for a particular type of mobile users and services. For example, a mobile user may choose to access a WLAN to send a large data file, but it may also select a 3G cellular network to carry on a voice call. In order to take advantage of this heterogeneity, we advocate that the appropriate solution consists of integrating, in an intelligent manner, the existing wireless systems in a way that users can obtain their services via the best available mobile network. However, implementing this type of integrated system yields many challenges in mobile handset design, wireless system discovery, terminal mobility, security and billing (Cavalcanti et al., 2005).

Among the well-known integration efforts available in the literature, the 3G wireless initiatives (i.e., 3GPP and 3GPP2) aim to integrate 3G and WLAN interworks. This integration is based on the loose and tight coupling scenarios (3GPP, 2004)(3GPP2, 2006). The inter-system roaming is based on bilateral service level agreements (SLAs). It is obvious that the SLA's approach is not appropriate when the number of integrated networks/operators is high. Additionally, operators are reticent to make their database available to other operators. Furthermore, the proposed solutions lack of efficient mobility schemes to ensure seamless roaming. The IP multimedia subsystem (IMS) is also proposed to provide several kinds of mobile services in UMTS to transparently connect mobile networks and the Internet (3GPP, 2003)(3GPP, 2005). Nevertheless, this solution is basically proposed to integrate new UMTS services and it is designed to support only SIP mobility. Moreover, the use of SIP-based mobility for UMTS networks may present significant handoff latency compared to network layer solutions (Banergiee at al., 2004).

As stated above, mobility management constitutes one of the crucial issues to be investigated for an eventual integration of metropolitan networks since it should ensure seamless roaming with QoS guarantee across different wireless technologies. Seamless roaming refers to the fact that MNs could perform handoffs with minimum disruption time, low packet loss, limited handoff blocking rate and minimal signaling cost. Moreover, QoS mapping between various mobile systems should be guaranteed. This need an efficient handoff preparation based on context awareness and network selection. Context awareness consists of maintaining a global view in terms of context information relevant to the integrated networks. The evident way for an MN to get context information is to be equipped with multiple air-interfaces and keep them on at all time. However, keeping all of the MN interfaces on consumes battery power which is not practical for mobile user energy autonomy. Concerning network selection, it aims to choose, for a given MN, a best network destination that respects its requirements in terms, for instance, of signal quality, bandwidth, monetary cost, etc,. It is clear that, in these circumstances, we need to consider a large number of context parameters rather than the traditional received signal strength (RSS).

Practically, an integrating architecture for metropolitan mobile networks must include the following characteristics:

- retain the best features of all individual integrated networks;
- guarantee that each user is consistently connected to the best available network;
- ensure a high level of security and privacy;
- be scalable and operate on existing infrastructures;
- provide seamless mobility;
- do not require high deployment costs.

In reality, it is difficult to respect all of the aforementioned goals and challenges. Hence, the cornerstone to remember while designing a new architecture for metropolitan mobile networks is to consider a tradeoff between all of these requirements in order to avoid a large number of drawbacks pertaining to the previous integration schemes.

This paper proposes a new Hybrid Interworking Architecture (HIA) for metropolitan mobility support. More specifically, the main objective of HIA consists of integrating any type of existing/future wireless systems while hiding their heterogeneity from each other. Additionally, HIA guarantees inter-system authentication and billing through an independent authority referred to as Interworking Cooperation Server (ICS). Then, we introduce a mobility scheme that takes into account context awareness and network selection to ensure seamless roaming through the integrated networks. Finally, an analytical model is developed to study the effectiveness of the proposed interworking solution.

The remainder of this paper is structured as follows: Section II describes related work and background. Section III introduces the proposed interworking architecture. Section IV outlines the proposed mobility scheme. Performance analyses and numerical results are presented in Section V. Finally, Section VI concludes the paper.

5.2 Related work and background

Based on the aforementioned requirements, we advocate that an appropriate interworking architecture for metropolitan mobile networks should be IP-based in order to integrate any types of mobile technologies. Moreover, it should ensure seamless roaming by supporting efficient mobility schemes. In this section, we present a brief overview of the interworking architectures as well as the commonly used mobility schemes available in the literature.

5.2.1 Existing interworking architectures

According to the 3G/WLAN interworking scenarios defined in (3GPP, 2004)(3GPP2, 2006), service continuity and seamless roaming provision are the most important aspects to be considered in such integration. To deal with these two crucial features, the 3GPP has proposed two interworking architectures called tight and loose coupling. In the tightly coupled scenario, the WLAN gateway is connected directly to the 3G gateway router (GGSN for UMTS and PDSN for CDMA2000). Accordingly, the WLAN is seen as an extension of the 3G wireless network and the mobile node (MN) has to implement both 3G and WLAN interfaces. This approach can be used to easily handle real-time traffic, ensure seamless mobility and guarantee QoS control. However, both technologies must be owned by the same operator and their respective architectures need adaptations. Moreover, 3G core network capacities are insufficient to accommodate the bulky WLAN data traffic, since the core network nodes are designed to handle circuit voice calls and short packets (Buddhikot et al., 2003).

In the loosely coupled scenario, both networks are connected via the Internet at the highest level of their respective networks. This approach allows independent deployment and traffic engineering. Hence, 3G carriers can benefit from other providers' WLAN without extensive capital investments. However, loose coupled schemes cannot support service continuity to other access network during handover, thus loose coupled architectures generate high latency and packet loss. Furthermore, the QoS provisioning depends on the Internet conditions.

Built on these basic interworking architectures, several integrating models are available in the literature. For instance, in (Wang et al., 2001) the authors propose a boundary location register (BLR) approach to integrate any two adjacent networks with partially overlapping areas. This approach lacks of scalability since a border gateway is needed for each pair of adjacent networks. Moreover, this architecture assumes the existence of bilateral service level agreements (SLAs) which is not suitable when the number of the integrated networks increases. In (Havinga et al., 2001), the authors introduced a new architecture that distinguishes signal from data traffics. This architecture is scalable but it needs a development of two networks called: *basic access network* and *common core network* which deal respectively with signal and data traffics. It is obvious that the deployment of such architecture requires high costs. The GSM association has proposed a backbone that uses the GPRS Roaming eXchange (GRX)(Inter-PLMN, 2003) to integrate GPRS networks belonging to different operators. However, this architecture is limited to only one technology (i.e., GPRS networks). An all-IP based architecture is proposed in (Akyildiz et al., 2005), it introduces two interworking units called NIA (network interworking agent) and IG (interworking gateway), to ensure the integration of different wireless systems around an IP backbone. However, this proposal uses only the received signal strength (RSS) to provide inter-system roaming which is no longer appropriate for 4G networks. Furthermore, it does not provide any network selection mechanism which may lead to wrong handoff decisions. An integrated architecture and a radio interface selection mechanism are introduced in (Buddhikot et al., 2003), but this solution did not take into account users requirements since handoff decision is based only on the RSS signal quality. Another, architecture is proposed in (Makaya et al., 2007). It introduces an interworking decision engine (IDE) to ensure seamless roaming between heterogeneous mobile networks. However, it is mainly based on the loose coupling architecture which means that QoS requirements are entirely depending on the Internet conditions.

5.2.2 Mobility management overview

Traditionally, mobility management is performed at the network layer due to the use of the Internet Protocol (IP) that allows routing packets between different technologies. However, mobility is recently experienced at different layers of the classical protocol stacks. In this subsection, we give a brief overview of the well-known mobility solutions.

A/ IP layer mobility

The very common way to ensure MNs roaming through heterogeneous technologies consists of using the IP layer mobility. In this category, Mobile IPv6 (MIPv6) (Johnson et al., 2004) is the most popular mechanism that allows mobile nodes to remain reachable in spite of their movements within IP-based mobile environments. However, MIPv6 has some drawbacks, such as high signaling overhead, packet loss and handoff latency, thereby causing real-time traffic deterioration which can be perceived by users (Pérez-Costa et al., 2003). These weaknesses lead to the investigation of other solutions designed to enhance MIPv6. The IETF proposed two main MIPv6 extensions: the Hierarchical MIPv6 (HMIPv6) (Soliman et al., 2005) and the Fast handover for MIPv6 (FMIPv6) (Koodli, 2005). These protocols tackle intra-domain or micro-mobility, while MIPv6 is used for inter-domain or macro-mobility. HMIPv6 handles handoffs locally through a special node called Mobility Anchor Point (MAP). On the other hand, FMIPv6 was proposed to reduce handoff latency and minimize service disruption during handoffs pertaining to MIPv6 operations, such as movement detections, binding updates and address configurations.

B/ Application layer mobility

Handling mobility at the application layer has also received a lot of attention since this category of solutions is almost independent of the underlying technologies. To accomplish this type of mobility, the SIP protocol (Handley et al., 1999) is widely used. Thus, when a mobile node moves during an active session into different networks, it first receives a new address, and then sends a new session invitation to its correspondent node. Subsequent data packets are forwarded to the MN using this new address. However, SIP by itself does not guarantee the maintenance of established Transmission Control Protocol (TCP) sessions or User Datagram Protocol (UDP) port bindings when moving, so further extensions such as S-SIP (Zhang et al., 2007) are needed to provide seamless handover capabilities.

C/ Transport layer mobility

In the last few years, transport layer-based mobility is gaining attention since it does not require a concept of home network and mobile node can perform smooth handovers if they are equipped with multiple interfaces. Moreover, this category of mobility schemes can benefit from flow control and the possibility to pause transmission during the handoff period. The first transport layer mobility solutions were based on TCP, and then other interesting mobility approaches have been proposed with the standardization of SCTP (Stewart, 2007) and mSCTP (Stewart et al., 2007).

- TCP-based mobility

Recently, several transport layer mobility schemes have been proposed to benefit from the connectivity facilities and flow control offered by the transport layer (Shaojian et al., 2004). From this perspective, a new TCP protocol architecture was proposed to support mobility (Hsieh et al., 2003). However, tremendous changes must be performed over the entire network to reach this goal. MSOCKS (Maltz et al., 1998) is another TCP-based proposal which does not require changes to the network layer infrastructure. However, it suffers from high latency and packet loss, since it follows a make-after-break approach (disable MN connections until a new path is ready). Migrate (Snoeren et al., 2000) is another TCP-based mobility solution which aims to ensure transparent TCP connection migration. Nevertheless, this solution requires changes to TCP implementation at both ends of the connection.

- SCTP-based mobility

Performing mobility at the transport layer has became more realistic with the emergence of the Stream Control Transmission Protocol (SCTP) (Stewart, 2007) and even more so with its mobile extension referred to as mSCTP (Stewart et al., 2007). Indeed, SCTP is a new transport layer protocol that was recently standardized under the RFC 4960. It inherited many TCP properties, but it also introduces novel and interesting features such as multistreaming and multihoming. Multistreaming consists of delivering independent data streams by decoupling reliable deliveries from message ordering. This feature prevents receiver head-of-line blocking in cases where multiple independent data streams occur during a single SCTP session (Scharf et al., 2006). On the other hand, multihoming allows an SCTP node to be reached through multiple IP addresses (interfaces). In fact, two SCTP nodes can exchange data by defining a common association. In SCTP terminology, an association is equivalent to a TCP connection. End points can be single-homed or multihomed. When single-homed, SCTP nodes are defined as [IP address: SCTP port], otherwise they are designated as [IP1 address, IP2 address…IPn address: SCTP port]. When establishing an association, end points define their primary path, as well as the secondary ones. The primary path is used to transfer data, while secondary paths are used for retransmissions and backups in the event of primary path failures. The SCTP ADDIP Extension enables SCTP nodes to dynamically add, delete and modify their primary address) without terminating an ongoing association.

In this sense, authors in (Ma et al., 2004) propose an approach to ensure vertical handoffs between UMTS and WLAN networks using SCTP multi-homing capabilities. In (Fu et al., 2004), a TraSH mobility scheme was proposed to perform seamless handovers between heterogeneous networks. In SIGMA (Fu et al., 2005), the authors propose an SCTP-based mobility architecture that integrates location management to ensure seamless handovers. In (Koh et al., 2004), the authors advance certain triggering rules to improve throughput during SCTP-based handoffs. All of these proposals are based on the Mobile SCTP extension (mSCTP) and their corresponding mobility procedure is summarized in Fig. 5.2.

1. ***Obtain an IP address from a new location.***
2. ***Add the new IP address to the association.***
3. ***Change the primary IP address.***
4. ***Delete the previous IP address from the SCTP association.***

Figure 5.2 Mobile SCTP-based handoff procedure

5.3 Proposed interworking architecture

To render more realistic the integration of heterogeneous wireless technologies within a metropolitan coverage and guarantee *always best connected* features to mobile users, this paper proposes a new integrating architecture referred to by Hybrid Interworking Architecture (HIA). Instead of developing new technologies and infrastructures, HIA aims to exploit and extend existing integration solutions in a way to make mobile user roaming more adapted to MAN's requirements in terms of mobility and heterogeneity support. For the sake of simplicity, only cdma 2000/3GPP2, UMTS/3GPP and WLAN networks are considered. In the rest of this paper, words: mobile node, mobile user and end user will be used interchangeably.

5.3.1 Hybrid Interworking Architecture

The proposed interworking architecture is illustrated in Fig. 5.3, it integrates different mobile networks around an IP backbone that hides their heterogeneities to each other. Each integrated network appears as a peer-system that can belong to independent operators. HIA is an open architecture that provides a coexistence platform for any number of mobile systems. For instance, it may integrate WLANs, 3G, WiMAX and ad hoc/sensor networks as it is shown in Fig. 5.3.

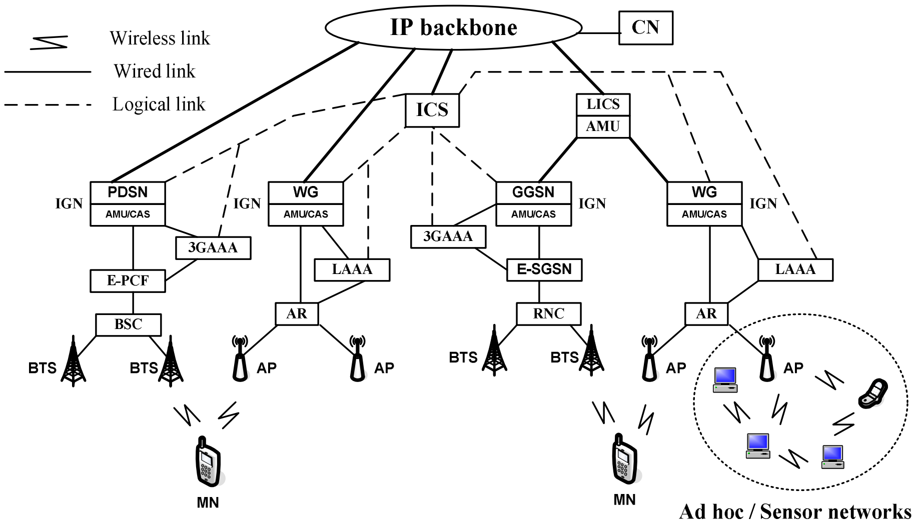


Figure 5.3 Hybrid Interworking Architecture (HIA) for MANs

To ensure interworking between heterogeneous networks, we introduce two novel entities designated by Interworking Cooperation Server (ICS) and Local Interworking Cooperation

Server (LICS). The ICS unit operates at the control plane since it manipulates only signaling traffic, while the LICS unit handles real traffic to guarantee seamless roaming to MNs running applications requiring high QoS conditions. ICS and LICS can be owned by an independent authority and could be seen as a value-added service that operators offer to their subscribers. Similar to the GPRS Roaming eXchange (GRX), the ICS mediates and manages service level agreement (SLA) between networks. Accordingly, an operator needs to establish only one direct SLA with the ICS instead of establishing individual SLAs with all other operators.

To guarantee service continuity and thus enable HIA with mobility features, the border nodes of the integrated networks are enhanced. For instance, the Serving GPRS (General Packet Radio Service) Support Node (SGSN) and Packet Control Function (PCF) are extended with access router functionalities and called respectively E-SGSN and E-PCF. Additionally, the Gateway GPRS Support Node (GGSN), the Packet Data Serving Node (PDSN) and the Wireless Gateway (WG) nodes are extended with interworking features and called Interworking Gateway Node (IGN). These features concern, especially, the AMU (Anchor Mobility Unit) and CAS (Context Aware Server) functionalities. More details concerning the IGN functionalities are provided in the next subsection. Furthermore, the GGSN, PDSN and WG components are endowed with router functionalities to perform message formats conversion and QoS mapping. Notice that all of the proposed enhancements are performed with existing entities which is expected to ensure HIA scalability at lower costs.

To perform authentication and billing when mobile users roam through heterogeneous technologies, the ICS cooperates with each local AAA (Authentication, Authorization and Accounting), ie., LAAAs (Local AAA) and 3GAAAs databases. This cooperation consists of granting MNs to access services that belongs to different operators, and provides a final billing report depending on the service charging policies observed in each one of the visited mobile systems. In what follows, we introduce the logical components pertaining respectively to the ICS, LICS and IGN components.

5.3.2 Interworking Cooperation Server (ICS)

The rationale of introducing the ICS is to ensure seamless roaming regardless of wireless technologies and service providers. In other words, the ICS is designed to coordinate information exchanged between heterogeneous wireless systems in order to reduce signaling load during mobile users roaming. More specifically, it coordinates the exchange of context information and mediates authentication between heterogeneous systems. Moreover, the ICS unit assists MNs while performing network selection and cooperates with local AAA components to ensure authentication and billing. The ICS can serve several operators since it manages only signaling traffic. However, if the number of networks and their respective subscribers increase, the ICS can be deployed in a hierarchical way. The logical components relevant to the ICS are presented in Fig. 5.4.

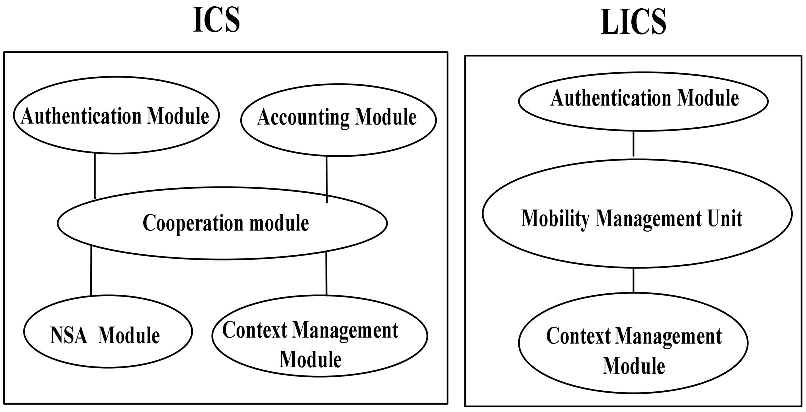


Figure 5.4 ICS and LICS logical components

The main role of the *Authentication Module* (*AuM*) is to perform mobile user authentication without extensive signaling costs. In fact, the introduction of the ICS unit avoids the use of direct security agreement between Home Networks (HN) and Foreign Networks (FN). Actually, bilateral security approach is not practical and presents high signaling costs when the number of integrated networks is important. All of the integrated networks through the proposed architecture have to make a registration entry with the ICS which provides them with a Network Security Agreement (NSA). By default, networks possessing an NSA are willing to share their services and their context information with the ICS. However, some exceptions may take place and must be reported to the ICS during the registration process. Locally, each integrated network provides its subscribers with a Security Passport (SP) which includes all of the necessary information concerning the NSA agreement. The SP's information is encrypted with the public key of the ICS. Hence, when an MN enters into an FN, it uses its SP for authentication. The FN's Local AAA verifies the SP token and decides accordingly to grant or not MN's requests. Notice that with this procedure, we do not need to invoke the ICS whenever we authenticate an MN.

The *Accounting Module* (*AccM*) aims to coordinate billing between different operators. Indeed, an MN can roam through heterogeneous networks whose charging policies may vary (connection, duration, transferred data, etc.). More specifically, when an MN is authorized by the FN, the local AAA unit (LAAA/3GAAA) maintains an Account Register Record (ARR) where information relevant to the current charging policies is stored and then, the ARR is sent to the ICS. Based on the HN billing policy, the ICS maps the MN's ARR to the format supported in the HN; then the charging information is forwarded to the MN's home network for billing purposes. Note that integrated networks must regularly update their charging policy with the ICS authority.

The *Cooperation Module* (*CoM*) cooperates with mobile users and mobile systems to ensure seamless roaming through heterogeneous technologies. Indeed, it provides QoS mapping between different service providers and coordinates context information exchanges with context aware servers (CAS) pertaining to each one of the integrated networks. Additionally, the *CoM* mediates between different services and network providers during the establishment of the Network Security Agreements. Furthermore, when MNs are subjected to network-controlled handoffs, the *CoM* assists mobile users to complete network selection by computing their handoff decision algorithms for example.

The *NSA Module* (*NsaM*) stores information pertaining to networks or operators having a security agreement entry with the ICS.

The rationale for the *Context Management Module* (*CMM*) is to manage context information pertaining to the integrated networks. This information may concern QoS parameters such as traffic status, average residence time, bandwidth, call blocking rate, etc. Practically, this module operates as a database that contains basic context information entries relevant to the integrated networks and has permanent exchanges with the *CoM* module.

5.3.3 Local Interworking Cooperation Server (LICS)

The LICS is a local interworking unit introduced to integrate heterogeneous networks in order to efficiently handle real-time traffic and ensure seamless mobility between them. More specifically, the LICS is enabled with packet redirection feature that allows traffic switching between different mobile systems. In addition, it converts high transmission rate to lower one and translates signaling message formats pertaining to heterogeneous wireless technologies. The logical components relevant to the LICS are shown in Fig. 5.4.

The main concern of the *Authentication Module*, relevant to the LICS, consists of authenticating mobile users and decides whether the MNs are granted to perform traffic redirection through the integrated networks. To complete the authentication task, the LICS receives periodic authentication updates from the ICS. These updates concern the list of network security agreements (NSA) recently registered at the ICS.

The *Mobility Management Unit* implements the Anchor Mobility Unit (AMU) functionalities which consist of performing traffic redirection between domains served by the same LICS unit. In addition, it converts high transmission rate to lower rate and translates signaling message formats between mobile systems that it integrates. Finally, the *Context Management Module* operates exactly like the one pertaining to the ICS. This means that it coordinates context information exchange between context aware servers belonging to mobile networks integrated through an LICS unit. Hence, it maintains a general view of context information relevant to the networks that it serves.

5.3.4 Interworking Gateway Node (IGN)

As stated earlier, the Gateway GPRS Support Node (GGSN), the Packet Data Serving Node (PDSN) and the Wireless Gateway (WG) nodes are extended with interworking features and called Interworking Gateway Node (IGN). These features concern, especially, the AMU (Anchor Mobility Unit) and CAS (Context Aware Server) functionalities.

Basically, the AMU functionalities consist of buffering traffic during the disruption period and performing redirection when the MN is attached to the new link. The AMU process is depicted in Fig. 5.5.

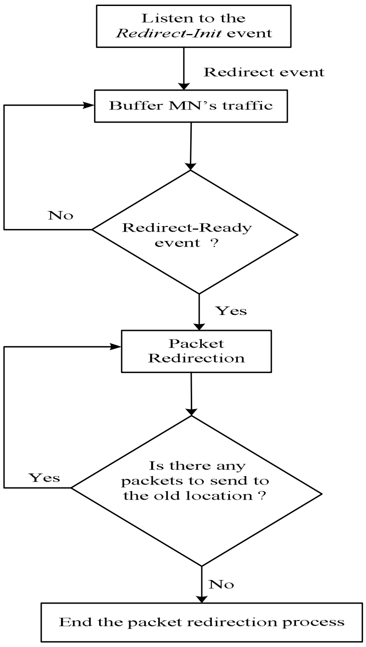


Figure 5.5 The AMU functionalities

More specifically, the AMU continuously listens to the redirect events (*Redirect-Init*). Once a *Redirect-Init* event occurs, the AMU starts buffering traffic sent to the old MN's IP address. When the MN is attached to its new location, it sends a *Redirect-Ready* message to notify the AMU that it is ready to receive data on its newly configured IP address. The AMU redirect process ends when no more packets are sent to the old MN address. The following section provides further details pertaining to the proposed handoff procedures when dealing with local and global mobility.

However, to execute an inter-system handoff, the MN has to select an appropriate network destination that respects its requirements. Thus, the MN should get, in advance, context information from its candidate neighbors. Practically, operators and private mobile networks are reticent to the idea of sharing their context information databases. An eventual possibility to get context information consists of using the Candidate Access Router Discovery protocol (CARD)

(Leibsh et al., 2005) which aims to reduce latency, packet loss and avoid the re-initiation of signaling from the beginning during a handoff. However, acquiring context information with the CARD protocol requires L2 ID detection which is possible only when the associated air-interfaces are always on. Additionally, authentication is needed between entities exchanging context parameters which yield addition delays and render the authentication procedure very difficult to execute when the number of the MN's neighbors increases. Thus, we propose to use local context aware servers that provide context information relevant to their serving home network.

More specifically, the Context Aware Server (CAS) aims to ensure context information gathering, by exchanging periodic context-beacons with ARs that it serves. Practically, access routers and access points (ARs/APs) are continuously aware about connection state of any MNs they serve. This task is achieved by gathering context information which may include: subnet status load, MN's pattern movement, channel number and frequencies, average residence time, etc,. Then, each ARs/APs sends periodic context-beacon to its serving IGN. The context-beacon message contains additional information such as the AR's prefix address, available bandwidth, traffic status, etc,. In this way, the IGN maintains a global view of all AR's domains that it serves. Additionally, the context-beacon messages are also sent both to the LICS (when it exists) and to the ICS. Hence, the LICS/ICS can also maintain a global view of the subnets belonging to the IGN that they serve.

Modules relevant to a context aware server (CAS) are illustrated in Fig. 5.6.

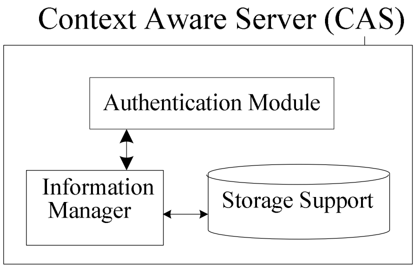


Figure 5.6 The CAS logical modules

- *The Authentication Module (AM)*

The *AM* refers to the entity that communicates with external components and authenticates mobile users.

- *The Information Manager (IM)*

The information manager (IM) manages local context information (inside a subnet or network) and sends periodic context-beacons to update CAS's profile at the LICS/ICS.

- *Storage Support (SS)*

The main role of this entity is to store local network context information. However, it can also manage basic operations such as deleting obsolete information and providing novel data structures for new ones.

5.4 Proposed handoff roaming scheme

The proposed interworking architecture is designed to integrate heterogeneous mobile systems located inside a metropolitan area. However, to render this integration realistic in terms of service continuity and QoS guarantee, we propose a new Hierarchical Transport layer Mobility scheme (HTM) that takes into account context awareness and network selection. We choose to experience mobility at the transport level for many reasons. First, transport layer based mobility do not use the concept of home and foreign networks and their relevant units such as HA (home agent) and FA (foreign network). Moreover, the transport layer offers flux control and the possibility to pause transmissions during the handoff period. Second, IP layer based mobility is not suitable for applications sensitive to QoS deterioration (Zeadally et al., 2007). Third, SIP by itself does not guarantee the maintenance of established sessions or user port bindings (Zhang et al., 2007). Fourth, the SCTP/mSCTP protocols offer some interesting features such as multihoming and dynamic address reconfiguration which is very useful to alleviate seamless roaming. Nevertheless, the SCTP based solutions available in the literature lack to deal with local mobility and guarantee QoS while roaming through homogeneous and heterogeneous systems. More specifically, the proposed mobility scheme is endowed with handoff preparation that takes into account context awareness and network selection.

5.4.1 Handoff preparation

Wireless technology has recently witnessed rapid progress in mobile user devices and network infrastructure. Thus, it is expected that mobile handsets will become more intelligent and sensitive to link layer changes. In other words, mobile terminals will be able to detect as quickly as possible handoff triggering events. With this new mobility features, we advocate that an appropriate handoff preparation will participate to avoid sever deterioration in QoS requirements and service continuity. In this subsection we present the different aspects that aim to enable end users with efficient preparation modules. These aspects include mobile node authentication, context information gathering and network selection.

Mobile nodes authentication

While roaming through a MAN, end users are expected to visit mobile networks that belong to different operators and support various technologies. Hence, to reduce signaling load due to the execution of the AAA procedure whenever an MN requests service/registration, we adopt the passport security based approach introduced in the previous section (section III) to authenticate mobile nodes. More specifically, when an MN enters into foreign network/domain, it uses its security passport (SP) for authentication. The foreign network/domain local AAA verifies the SP token and decides accordingly to grant or not MN's requests (i.e., service or registration). Recall, that the SP is provided by the MN's home network (HN) based on the NSA information that the HN had obtained while completing its first registration entry with the ICS.

Network selection

When an MN receives a trigger event (TE), i.e., weak signal strength, poor bandwidth, high traffic status, etc., it sends a *NDS_Req* (New Destination Selection Request) message to its serving IGN. The *NDS_Req* message includes information such as user preferences (weights) and thresholds (minimum QoS). Upon receiving the *NDS_Req* message, the IGN verifies whether the MN can perform an L2 handoff. When this option is possible (e.g., performing an L2 handoff), IGN sends a *NDS_Rep* reply that contains the prefix address of the selected NAR (new access router). Otherwise, the *NDS_Req* message is sent to its serving LICS/ICS. Once this message is received, the LICS/ICS sends a *Context_Get_Infos* message to the entire CASs located in the MN's vicinity. Then, each CAS replies with a *Context_Get_Rep* that contains the requested information. After that, the LICS/ICS computes a pre-configured preference score function and provides a list of candidate ARs (resp networks) that respects user requirements. Recall that this kind of pre-configured function can be defined during the system setup. Then it uses the *NDS_Rep* message to send the list of candidate destinations that satisfy MN's requirements.

Once, the MN receives its *NDS_Rep* message, it selects one AR/network destination and turns on its associated wireless interface. Then, it performs authentication and obtains a new IP address from the selected destination by using the IPv6 auto-configuration (Thomson et al., 1998) or the DHCP (Droms, 1997) features. At this stage, the MN is ready to perform either horizontal or vertical handoff.

Fig. 5.7 depicts the main steps of the proposed handoff preparation approach.

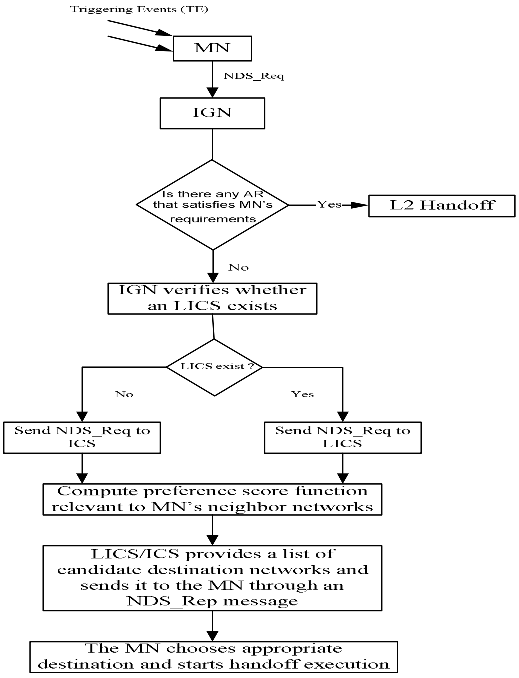

Triggering Events (TE)
MN
NDS_Req
IGN
Is there any AR that satisfies MN's requirements
Yes
L2 Handoff
No
IGN verifies whether an LICS exists
LICS exist ?
No
Yes
Send NDS_Req to ICS
Send NDS_Req to LICS
Compute preference score function relevant to MN's neighbor networks
LICS/ICS provides a list of candidate destination networks and sends it to the MN through an NDS_Rep message
The MN chooses appropriate destination and starts handoff execution

Figure 5.7 Handoff preparation process

Notice that with this approach, the MN receives only a list of its candidate destinations rather than handling their context information. Moreover, this approach reduces signaling traffic in the wireless link since it avoids requesting individually all of the MN's neighbors for context information. Furthermore, delays relevant to the authentication of the MN with each one of its neighbor networks, are avoided because the requested context information is obtained through the ICS/LICS which are assumed to have a secure entries with mobile systems located at the MN's vicinity.

5.4.2 Handoff execution

As stated earlier, the realistic transport layer mobility schemes are based on SCTP and especially on its mobile version mSCTP. Nevertheless, the previous efforts do not take into account QoS guarantee and indure unnecessary handoff delays when an MN roams inside a same mobile technology. Hence, we introduce a new Hierarchical Transport layer Mobility scheme (HTM) that runs over the proposed architecture. Thus, in the rest of this paper, this proposal is noted as *HTM/HIA*. Practically, *HTM/HIA* can be seen as a cross layer mobility scheme that uses Layer 2 for handoff triggering events (TE), Layer 3 for address configuration as well as traffic redirection (i.e., through the AMU unit) and uses transport layer (SCTP/mSCTP) to achieve end-to-end mobility support. To illustrate how the proposed *HTM/HIA* operates, consider the roaming scenario depicted in Fig. 5.8. We notice that in the presence of IGN/LICS units, an MN can perform either handoff of type (a), (b). Otherwise, it performs a handoff of type (c).

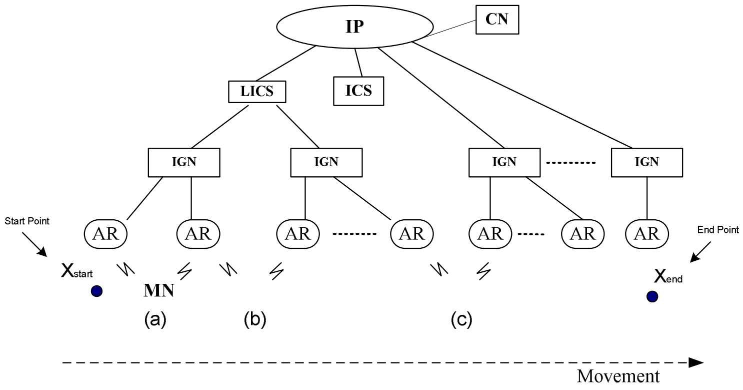


Figure 5.8 MN roaming topology

In order to consider local mobility (handoff of type (a)) as well as the inter-system handoffs performed in the presence of an LICS unit (handoff of type (b)), we introduce an AMU whose main role consists of assisting mobile nodes to perform seamless handoffs as it has been already introduced in the AMU process outlined in Fig. 5.5.

C/ Handoff Procedures

In the last few years, mobile devices are becoming increasingly powerful, intelligent and sensitive to link changes. Thus, it can be assumed that an MN detects its movement toward a new access router by using L2 triggers such as weak signal strength, high bit error rate, etc. Hence, when an MN receives an L2 triggering event, it sends a *NDS_Req* (New Destination Selection Request) message to its serving IGN to initiate a network selection process. According to the received *NDS_Rep*, the MN may perform either an HTM / HIA^{local} procedure (local handoff) or an HTM / HIA^{global} procedure (global handoff).

- Local Handoff Procedure (HTM / HIA^{local})

The HTM / HIA^{local} procedure is initiated when an MN perform a handoff between subnets served by a same IGN/LICS components. In this case, it obtains an IP address from the selected destination through its serving IGN/LICS unit. Practically, this task can be completed through a router solicitation message (RAS) by using either DHCP (Droms, 1997) or IPv6 autoconfiguration (Thomson et al., 1998). The IGN/LICS keeps an association entry between the new obtained address and the one currently used by the MN. From this time, the MN is ready to initiate a local handoff process. Recall, that until now the MN continues to receive data on its old path. When the MN decides to move to its new location, it sends a *Redirect-Init* message to the IGN/LICS unit. This message informs the IGN/LICS that the MN is performing an L2 link switching (L2 handoff). At this time, the IGN/LICS starts buffering all the packets sent to the MN's previous address until the MN attaches to its new access router (NAR) link. As soon as the MN is attached to the NAR, it sends a *Redirect-ready* message to notify the IGN/LICS that it has been successfully attached to its new location. Upon receiving the *Redirect-ready* message, the IGN/LICS starts packet forwarding to the new MN's IP address. At the same time, the MN sends an *ADDIP_Soft* chunk to inform its correspondent node (CN) that a handoff had occurred and it has to set the new MN's IP address as the primary path of their association. Finally, when the MN is completely far from its previous attachment point, the old path is deleted. The entire $HTM\ /\ HIA^{local}$ procedure is illustrated in Fig. 5.9.

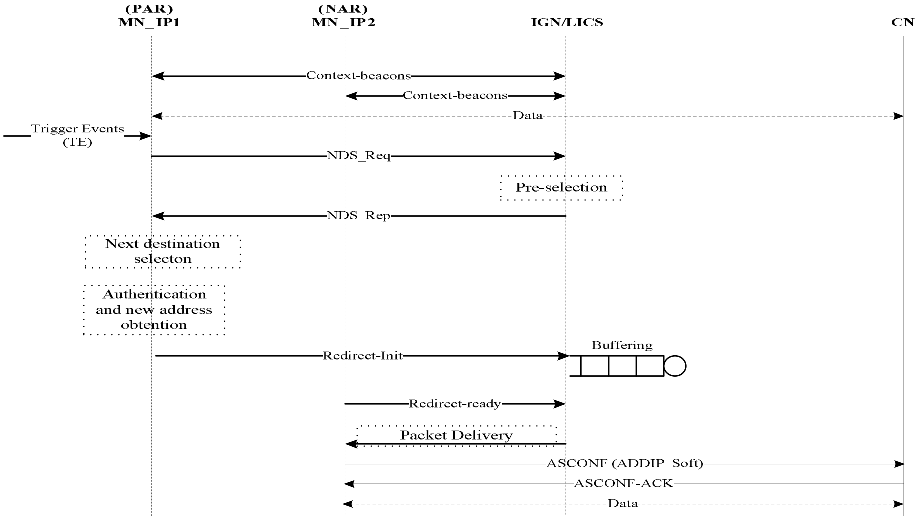


Figure 5.9 HTM/HIA^{local} handoff procedure

ADDIP_Soft is a new chunk introduced to set the primary path when the MN is subjected to a local handoff. When the CN receives the *ADDIP_Soft* chunk, it concludes that its pair (MN) has performed a local handoff. The CN immediately transmits packets through the MN's new IP address (IP2) and ignores the previous one (IP1). The description of the new proposed *ADDIP_Soft* chunk appears in Fig. 5.10.

Type = 0xC008	Length = 20
Chunk-ID = 0x11122233	
Value = 0x0a010101 (New address)	
Value = 0x0a010111 (Old address)	

Figure 5.10 *ADDIP_Soft* chunk description

- Global Handoff Procedure ($HTM\ /\ HIA^{global}$)

When it is not possible to perform handovers through IGN/LICS units, all handoffs are completed with the $HTM\ /\ HIA^{global}$ procedure. This means that an MN, that wants to perform a handoff between subnets (or cells) served by different IGN/LICS, uses the $HTM\ /\ HIA^{global}$ handover procedure depicted in Fig. 5.11. More specifically, the execution of the handoff preparation process allows the MN to choose an appropriate network destination that satisfies its preferences. Then, it runs on the corresponding wireless interface and obtains a new address by using either DHCP or IPv6 autoconfiguration After that, the obtained address is dynamically added to the MN's active association by using the mSCTP extension. Recall that the MN continues receiving data on its current attachment point while completing these configurations. Once the MN decides to hand off into its new location, it sends an ASCONF (Set Primary Address) to the CN. From this time, the old path falls down and the new one becomes operational as soon as the MN receives an ASCONF-ACK chunk on its new IP primary address. When the MN is far from its old location, the old path is removed from the association.

The signaling messages relevant to the HTM/HIA^{global} procedure are depicted in Fig 5.11.

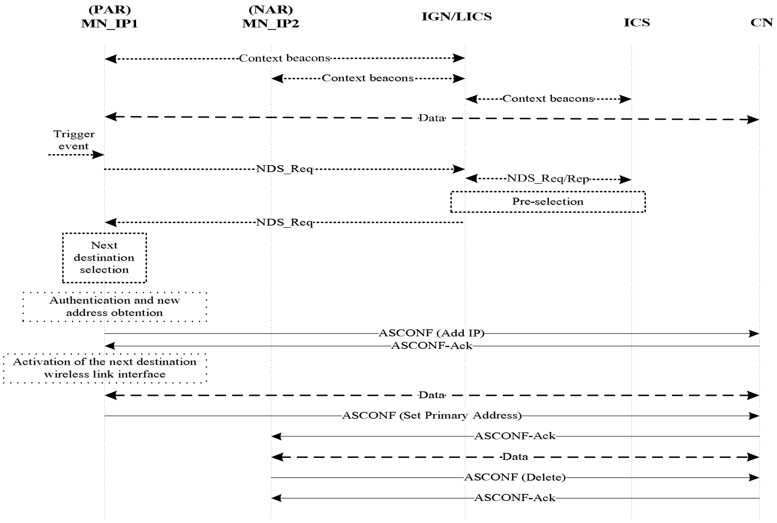


Figure 5.11 HTM/HIA^{global} handoff procedure

5.5 Analytical model

To study the impact of the proposed interworking architecture (HIA) on mobile node roaming inside an MAN (Metropolitan Area Network), we develop an analytical model to compare mSCTP based mobility with our *HTM/HIA* proposal. In fact, without the AMU unit (e.g., IGN/LICS), *HTM/HIA* turns into mSCTP. Thus, we advocate that such comparison will reflect how the proposed HIA improves mobile nodes roaming. Moreover, previous studies such as the ones introduced in (Fu et al., 2005) (Zeadally et al., 2007) advocate that SCTP/mSCTP based mobility is more appropriate for applications and services which are sensitive to QoS deterioration compared MIPv6 based solutions. That is why we focus our performance analyses on the mSCTP and *HTM/HIA* comparisons. In the rest of this paper, mSCTP refers to the handoff mobility procedure outlined in Fig. 5.2.

5.5.1 Preliminary and notations

As it is illustrated in Fig 5.8, an MN can perform one of the three handoff types referred to as (a), (b) and (c).

Where,

(a) : refers to handoffs between two access routers (AR) belonging to a same IGN domain,

(b) : refers to handoffs between two ARs belonging to different IGN domains served by the same LICS unit,

(c) : refers to handoffs between ARs belonging to different IGN domains without LICS units.

Let μ_r be the border crossing rate of an MN through access routers (ARs),

Let μ_d be the border crossing rate of an MN through IGN domains, Let μ_I be the border crossing rate through ARs belonging to a same IGN domain, where μ_I is defined as: $\mu_I = \mu_r - \mu_d$. According to (Bauman et al., 1994), if we assume that an IGN coverage area is composed of *M* circular access router subnets, the border crossing rates can be expressed as:

$$\begin{cases} \mu_d = \dfrac{\mu_r}{\sqrt{M}} \\ \mu_I = \mu_r \cdot \dfrac{\sqrt{M}-1}{\sqrt{M}} \end{cases} \qquad (1)$$

In the following analysis, we assume that mobile nodes roam under the fluid-flow mobility model introduced in (Wang et al., 2000). Thus, μ_r can be defined as: $\frac{\rho \cdot v \cdot R_s}{\pi}$, where: ρ is the user density, v the MN average velocity and R_s the perimeter of a subnet.

Let ε be the probability to perform a handoff in the presence of an LICS unit when an MN roams from X_{start} to X_{end}, i.e., percentage of handoffs completed in case (b) as shown in Fig. 5.8. It is evident that ε depends on the number of LICS units as well as on the number of IGN they serve. We define ε as:

$$\varepsilon = \frac{N_{IGN}^{LICS} - N_{LICS}}{N_{AR} - 1} \qquad (2)$$

Where:

$0 \le \varepsilon \le 1$,

N_{AR} : number of the overall AR domains, $N_{AR} \ge 2$,

N_{LICS} : number of LICS,

N_{IGN}^{LICS} : number of IGN domains served by an LICS.

We assume that each LICS serves at least one IGN in order to be sure that $N_{AMU}^{LICS} \ge N_{LICS}$.

Accordingly, if we denote μ_L as the border crossing rate relevant to handoff of type (b), μ_L will be defined as:

$$\mu_L = \varepsilon \cdot \mu_d = \frac{\varepsilon \cdot \mu_r}{\sqrt{M}} \qquad (3)$$

In order to study the effectiveness of the proposed mobility scheme we consider a traffic model composed of two levels, a session and packet. The MN mobility will be modeled by the cell residence time and a number of random values introduced in (Fang, et 2003). Generally, we model the incoming sessions as a Poisson process (i.e., inter-session arrival time are exponentially distributed). According to (Fang, et 2003), the inter-session arrival time may not be exponentially distributed. Thus, alternative distribution models such as Hyper-Erlang, Gamma and Pareto have been proposed. However, performance analyses show that the exponential approximation remains an acceptable tradeoff between complexity and accuracy (Fang, et 2003). Therefore, for simplicity we assume that the MN residence time in an AR subnet and in an IGN domain follow exponential distribution with parameters μ_r and μ_d respectively, while session arrival process follows a Poisson distribution with rate λ_s. Hence, if we denote: $E(N_r)$ as the average number of AR subnet crossing, $E(N_d)$ as the average number of IGN domain crossing and $E(N_I)$ as the average number of AR subnet crossing performed inside an IGN domain, we can define the above averages as introduced in (Xiao et al., 2004) by:

$$E(N_r) = \frac{\mu_r}{\lambda_s} \qquad (4)$$

$$E(N_d) = \frac{\mu_d}{\lambda_s} \qquad (5)$$

$$E(N_I) = \frac{\mu_I}{\lambda_s} \qquad (6)$$

Hence, the average number of IGN domain crossing in the presence of an LICS unit (i.e., handoffs of type (b) as mentioned in Fig. 5.8) is given by:

$$E(N_L) = \varepsilon \cdot E(N_d) = \varepsilon \cdot \frac{\mu_d}{\lambda_s} \qquad (7)$$

Similarly, the average number of the IGN domain crossing without (in the absence of) an LICS unit is given by:

$$E(N_{nL}) = (1-\varepsilon) \cdot E(N_d) = (1-\varepsilon) \cdot \frac{\mu_d}{\lambda_s} \qquad (8)$$

The notation used in our analysis is summarized in Table 5.1

Table 5.1 Notation

$T_{X,Y}$	transmission cost between node *X* and node *Y*
P_Z	processing cost at node *Z*
$N_{hop}^{X,Y}$	number of hops between node *X* and *Y*
δ	a proportionality constant to illustrate that the transmission cost for wireless hops are superior to those of wired hops
T_{hop}^{c}	transmission cost per hop
l_X	one lookup cost at node X
η_X	packet tunneling cost at node X
$D_{X,Y}$	transmission delay between nodes *X* and *Y*
$D_{tunneling}$	packet tunneling time
P_Z^t	processing time at node *Z*
T_{MD}	Movement Detection delay
T_{AC}	Address Configuration delay
T_{L2}	L2 handoff delay
T_{UF}	AMU Update and packet Forwarding delay

In what follows, we use the above equations to analyze both signaling and packet delivery costs of the studied mobility schemes.

5.5.2 Total cost analysis

We define the total cost (C_{total}) as the sum of signaling and packet delivery costs. In other words, C_{total} is given by:

$$C_{total} = C_{signal} + C_{delivery} \quad (9)$$

The signaling cost refers to the amount of signaling traffic while the packet delivery cost refers to the network overhead. Note that signaling cost required for L2 handoff and address configuration are not considered in our analysis since they are the same for the compared protocols.

5.5.2.1 *HTM/HIA* total cost

The *HTM/HIA* total cost is defined as:

$$C_{total}^{HTM/HIA} = C_{signal}^{HTM/HIA} + C_{delivery}^{HTM/HIA} \quad (10)$$

- *HTM/HIA* signaling cost

The *HTM/HIA* signaling cost is incurred when an MN performs either (a), (b) or (c) handoffs is given by:

$$C_{signal}^{HTM/HIA} = E(N_I) \cdot C^{AR} + E(N_L) \cdot C^{LICS} + E(N_{nL}) \cdot C^{IGN} \quad (11)$$

Where :

C^{AR} : refers to the signaling cost when an MN performs a handoff of type (a)

C^{LICS} : refers to the signaling cost when an MN performs a handoff of type (b)

C^{IGN} : refers to the signaling cost when an MN performs a handoff of type (c)

Moreover, if we assume that a handoff preparation is always followed by a handoff execution, the expressions relevant to C^{AR}, C^{LICS} and C^{IGN} are given in Table 5.2.

Table 5.2 Expression of signaling costs

C^{AR}	=	$T_{MN_p,IGN} + T_{MN_n,IGN} + 2 \cdot T_{MN_n,CN} + 2 \cdot P_{IGN} + P_{CN}$
C^{LICS}	=	$T_{MN_p,LICS} + T_{MN_n,LICS} + 2 \cdot T_{MN_n,CN} + P_{IGN} + 2 \cdot P_{LICS} + P_{CN}$
C^{IGN}	=	$3 \cdot T_{MN_p,CN} + 3 \cdot T_{MN_n,CN} + 3 \cdot P_{CN}$

Where MN_p and MN_n refer respectively to the MN's location before and after a handoff. The $T_{X,Y}$ cost can be expressed as:

$$T_{X,Y} = (N_{hop}^{X,Y} - 1 + \delta) \cdot T_{hop}^{c} \quad (12)$$

To illustrate the impact of the MN mobility and the MN average session arrival on the *HTM/HIA* signaling cost, we introduce a session-to-mobility factor (SMR) which represents the relative ratio of session arrival rate to the mobility rate.

The SMR factor is expressed by : $SMR = \frac{\lambda_s}{\mu_r}$ (13).

Hence, if we consider equations (1), (6), (7), (8) and (13), the equation (11) becomes:

$$C_{signal}^{HTM/HIA} = \frac{1}{SMR\sqrt{M}}\left[(\sqrt{M} - 1)C^{AR} + \varepsilon \cdot C^{LICS} + (1-\varepsilon) \cdot C^{IGN}\right] \quad (14)$$

- *HTM/HIA* packet delivery cost

Let A_p be the average packet sent by the CN during one session lifetime. Based on Fig. 5.8, the MN can perform either handoffs of type (a), (b) or (c). However, only handoffs of type (a) and (b) incur a table lookup and an IP tunneling costs at the IGN/LICS. Hence the *HTM/HIA* packet delivery cost is given by :

$$C_{delivery}^{HTM/HIA} = A_p \cdot T_{MN,CN} + E(N_I) \cdot (l_{IGN} + \eta_{IGN}) \cdot A_p^{(a)} + E(N_L) \cdot (l_{LICS} + \eta_{LICS}) \cdot A_p^{(b)} \quad (15)$$

5.5.2.2. mSCTP total cost

The mSCTP total cost is defined as:

$$C_{total}^{mSCTP} = C_{signal}^{mSCTP} + C_{delivery}^{mSCTP} \quad (16)$$

- mSCTP signaling cost

Based on the mSCTP handoff procedure depicted in Fig. 5.3, the mSCTP signaling cost is given by:

$$C^{mSCTP}_{signal} = (E(N_I) + E(N_L) + E(N_{nL})) \cdot (3 \cdot T_{MN_p,CN} + 3 \cdot T_{MN_n,CN} + 3 \cdot P_{CN}) \quad (17)$$

To express equation (17) as a function of the SMR factor, we use equations (1), (6), (7), (8) and (13).

$$C^{mSCTP}_{signal} = \frac{3}{SMR} \cdot (T_{MN_p,CN} + T_{MN_n,CN} + P_{CN}) \quad (18)$$

- mSCTP packet delivery cost

Since the mSCTP handoff procedure did not incur any IP tunneling or table lookup costs, its packet delivery is given by:

$$C^{mSCTP}_{delivery} = A_p \cdot T_{MN,CN} \quad (19)$$

5.5.3 Handoff Latency and Packet Loss

The handoff latency is defined as the time elapsed between sending the last data packet through the old MN's primary address (i.e., old location) and receiving the first data packet on the MN's new primary address (i.e., new location). The packet loss refers to the amount of packets lost during this disruption time.

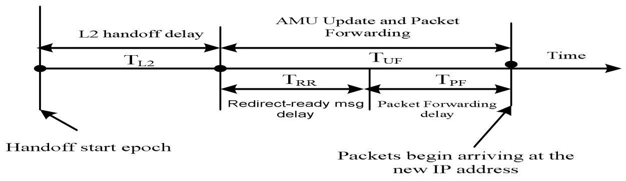


Figure 5.12 Timeline delay of *HTM/HIA* for intra IGN handoffs (type (a))

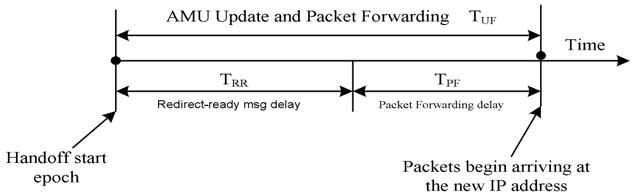

AMU Update and Packet Forwarding T_UF
Time
T_RR
T_PF
Redirect-ready msg delay
Packet Forwarding delay
Handoff start epoch
Packets begin arriving at the new IP address

Figure 5.13 Timeline delay of *HTM/HIA* for intra LICS handoffs (type (b))

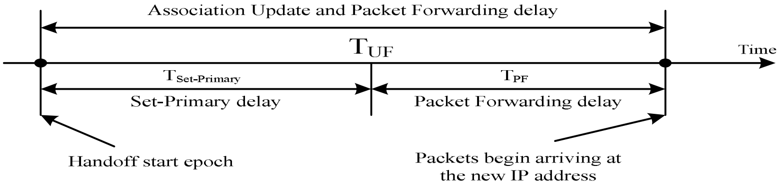

Association Update and Packet Forwarding delay
TUF
Time
TSet-Primary
TPF
Set-Primary delay
Packet Forwarding delay
Handoff start epoch
Packets begin arriving at the new IP address

Figure 5.14 Timeline delay of mSCTP and *HTM/HIA* vertical handoffs (type (c))

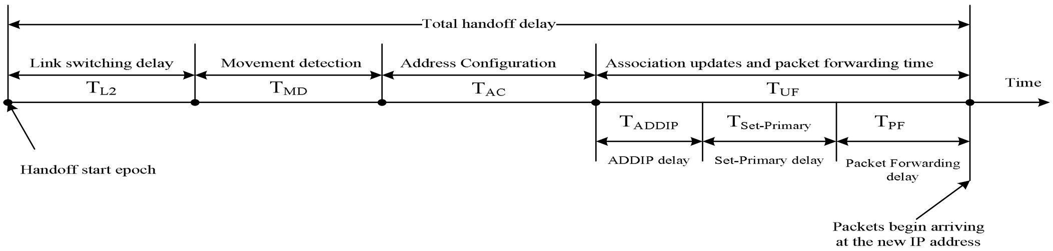


Figure 5.15 Timeline delay of mSCTP horizontal handoffs

If a mobile node moves between cells belonging to a same technology (horizontal handoff), it cannot simultaneously use its two interfaces (Atallah et al., 2006). However, if it performs a handover between heterogeneous wireless technologies (i,e., vertical handoff), it can use its wireless interfaces in parallel. This means, that the MN continues to receive traffic on its old path while it performs L2 link switching, movement detection, address configuration through the new interface and the association update (ADDIP). Practically, we can divide handoff latency into: link switching or L2 handoff delay (T_{L2}), movement detection delay (T_{MD}), address configuration delay (T_{AC}) and association updates and packet forwarding time (T_{UF}).

According to the handoff scenarios depicted in Fig. 5.8., an MN can perform either handoffs of type (a), (b) or (c). Hence, we define the average handoff latency for *HTM/HIA* as:

$$D^{HTM/HIA}_{handoff} = \frac{1}{E(N_I)+E(N_L)+E(N_{nL})}\cdot\left[E(N_I)\cdot D^{(a)}_{handoff}+E(N_L)\cdot D^{(b)}_{handoff}+E(N_{nL})\cdot D^{(c)}_{handoff}\right] \quad (20)$$

Where $D^{(a)}_{handoff}$, $D^{(b)}_{handoff}$ and $D^{(c)}_{handoff}$ refer respectively to the handover delay relevant to handoff types (a), (b) and (c). Based on the timing diagrams relevant to each one of the *HTM/HIA*'s handoff types illustrated in Fig. 5.12, Fig. 5.13 and Fig. 5.14, the corresponding expressions of $D_{handoff}^{(a)}$, $D_{handoff}^{(b)}$ and $D_{handoff}^{(c)}$ are given in Table 5.3.

Table 5.3 Expressions of *HTM/HIA* handoff delays

$D_{handoff}^{(a)}$	=	$T_{L2} + 2 \cdot D_{MN,IGN} + D_{tunneling} + P_{IGN}^{t} + \tau$
$D_{handoff}^{(b)}$	=	$2 \cdot D_{MN,LICS} + D_{tunneling} + P_{LICS}^{t} + \tau$
$D_{handoff}^{(c)}$	=	$2 \cdot D_{MN,CN} + P_{CN}^{t} + \tau$

If we consider equations: (6), (7) and (8), the relation (20) can be expressed as:

$$D_{handoff}^{HTM/HIA} = \frac{1}{\sqrt{M}} \cdot \left[(\sqrt{M} - 1) \cdot D_{handoff}^{(a)} + \varepsilon \cdot D_{handoff}^{(b)} + (1-\varepsilon) \cdot D_{handoff}^{(c)} \right] \qquad (21)$$

$D_{X,Y}$ is defined as:

$$D_{X,Y} = \frac{1-q}{1+q} \cdot \left(\frac{s}{B_{wl}} + L_{wl}\right) + (N_{hop}^{X,Y} - 1) \cdot \left(\frac{s}{B_w} + L_w + \varpi_q\right) \qquad (22)$$

Where s is the message size, ϖ_q is the average queuing delay at each intermediate router, q is the probability of wireless link failure, B_{wl} (resp B_w) the bandwidth of wireless (resp wired) link and L_{wl} (resp L_w) wireless (resp wired) link delay (McNair et al., 2001).

Similarly, the average mSCTP handoff latency is given by:

$$D_{handoff}^{mSCTP} = \frac{1}{E(N_I) + E(N_L) + E(N_{nL})} \cdot \left[E(N_I) \cdot D_{handoff}^{mSCTP,horizontal} + (E(N_L) + E(N_{nL})) \cdot D_{handoff}^{mSCTP,vertical} \right] \ (23)$$

Table 5.4 Expressions of mSCTP handoff delays

$D_{handoff}^{mSCTP,horizontal}$	=	$T_{AC} + T_{MD} + T_{L2} + 4 \cdot D_{MN,CN} + P_{CN}^{t} + \tau$
$D_{handoff}^{mSCTP,vertical}$	=	$2 \cdot D_{MN,CN} + P_{CN}^{t} + \tau$

On the other hand packet loss is proportional to the handoff delay since all data packets exchanged during this disruption period are lost. Practically, the packet loss is defined for both *HTM/HIA* and mSCTP as:

$$\begin{cases} P_{loss}^{HTM/HIA} = \lambda_p \cdot D_{handoff}^{HTM/HIA} - Min(B_{HTM/HIA}, B_{IGN/LICS}) \\ P_{loss}^{mSCTP} = \lambda_p \cdot D_{handoff}^{mSCTP} \end{cases} \quad (24)$$

Where $B_{HTM/HIA}$ refers to the buffer size required for *HTM/HIA* and $B_{IGN/LICS}$ is the buffer size available at the IGN/LICS unit. The buffer size required for *HTM/HIA* is proportional to packet arrival rate. This buffer is computed for intra IGN handoffs (type (a)) as:

$$B_{HTM/HIA} = \lambda_p \cdot (T_{L2} + T_{UF}) \ (25)$$

In the case of intra LICS and vertical handoffs (i.e. handoffs of type (b) and (c)), the $B_{HTM/HIA}$ buffer is estimated as :

$$B_{HTM/HIA} = \lambda_p \cdot T_{UF} \ (26)$$

In other words, $B_{HTM/HIA}$ refers to the buffer size estimated while considering the *HTM/HIA* timeline diagram depicted respectively in Fig 5.12, Fig 5.13 and Fig 5.14.

5.5.4 Handoff blocking probability

The handoff blocking probability ($P_{blocking}$) refers to the fact that an ongoing session will be terminated prematurely due to unsuccessful handoff during a session lifetime. This factor is very important since mobile users are more sensitive to call disruption during a session than when the call is initiated. $P_{blocking}$ is defined by:

$P_{blocking} = P_{prob}(t_{handoff} > t_s)$, where $t_{handoff}$ is the random variable defining the handover period and t_s designates the average subnet residence time. If we assume that $t_{handoff}$ is exponentially distributed with $F_T(t)$ as a density function, $P_{blocking}$ can be expressed as:

$$P_{blocking} = \int_0^{\infty} [1 - F_T(x)] f_s(x) dx = \frac{\mu_r \cdot t_{handoff}^{mean}}{1 + \mu_r \cdot t_{handoff}^{mean}} \qquad (27)$$

Where: $t_{handoff}^{mean}$ refers to the mean value of the total handoff latency.

5.5.5 Processing Load of the ICS

To evaluate the charge incurred at the ICS when MNs roam through heterogeneous networks, we propose to compare this processing load to the one generated at the HA (home agent) by a similar number of handovers. Based on the roaming scenario depicted in Fig. 5.8., an MN can perform either handoffs of type (a), (b) or (c). With MIPv6, all of these handoffs (i.e., (a), (b) and (c)) incur a binding update with the HA. Thus, if we denote P_{HA} as the processing binding time at the HA, the corresponding load is given by:

$$L_{HA} = E(N_r) \cdot N_{MN} \cdot P_{HA} = \frac{\mu_r}{\lambda_s} \cdot N_{MN} \cdot P_{HA} \qquad (28)$$

Where N_{MN} designates the average number of mobile nodes present throughout the integrated networks. On the other hand, with *HTM/HIA*, the binding updates are performed locally when the MNs roams inside the same IGN domain or between IGN domains served by the same LICS unit. Otherwise, the ICS is invoked during the handoff preparation phase as illustrated in Fig. 5.7. Let P_{ICS} denotes the processing time at the ICS; the processing load relevant to the IGN average crossing rate is given by:

$$L_{ICS} = E(N_{nL}) \cdot N_{MN} \cdot P_{ICS} = \frac{(1-\varepsilon)}{\lambda_s \cdot \sqrt{M}} \cdot \mu_r \cdot N_{MN} \cdot P_{ICS} \qquad (29)$$

Hence,

$$\frac{L_{ICS}}{L_{HA}} = \frac{(1-\varepsilon)}{\sqrt{M}} \cdot \frac{P_{ICS}}{P_{HA}} \qquad (30)$$

Accordingly if we assume that both the ICS and HA are equipped with high computing capabilities, we can consider that $P_{ICS} \cong P_{HA}$. Therefore, $L_{ICS} \leq L_{HA}$ since $M \geq 2$ and $\varepsilon \leq 1$.

5.6 Performance evaluation

In this section we present results relevant to the conducted comparisons based on both simulation and numerical results. We choose mSCTP as the benchmark transport layer mobility protocol for our comparison since all the previous SCTP-based mobility proposals use the mSCTP standard. Moreover, mSCTP based mobility is considered as an interesting alternative especially for applications with high QoS requirements (Zeadally et al., 2007).

5.6.1 Simulation setup

The main concern of our simulations is to show how the introduced IGN/LICS unit improves handoff seamlessness. That is why we consider the simulation scenario depicted in Fig. 5.16. This scenario is designed in such a way to provide realistic results, while remaining sufficiently small to be handled efficiently with the ns-2 simulator. Simulation code is based on the SCTP module developed at the University of Delaware. This SCTP module is modified so that it can support the newly introduced ADDIP-Soft Chunks, as well as AMU functionalities.

Initially, the MN is assigned to AR1 and benefits from an ongoing association with CN. When the MN moves from AR1 to AR2, it performs a local handoff (inside an AMU). In all simulations, the observed MN moves at various speeds, on a straight line, from AR1 to AR2 sub-network. Each AR operates according to the 802.11b (11 Mbit/s) standards in the Distributed Coordination Function (DCF). Delays for both 802.11b WLANs equal 15 ms. A CBR agent is attached to either CN or MN depending on the metric to be measured (i,e., latency or throughput). The average experiment time lasts around 300 s.

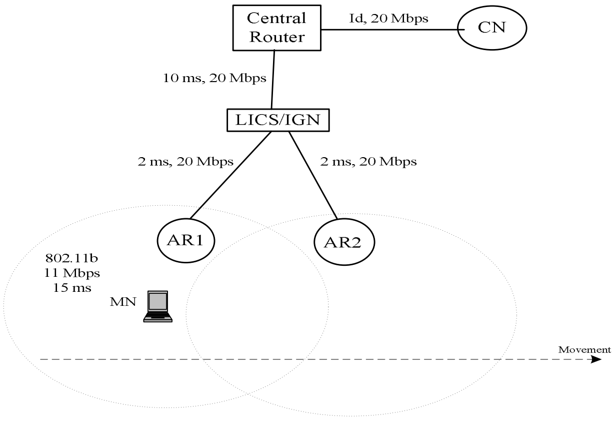


Figure 5.16 Simulation network topology

5.6.2 Simulation results

Fig. 5.17 illustrates handoff latency behavior when an MN completes HTM / HIA^{local} and mSCTP handoffs. In fact, several experiments were conducted where the MN performs a handoff from AR1 to AR2, then it returns back to AR1. In each experiment, a wired hop is added between the MN and the CN, meaning that an additional delay is added to the CN-AMU link. The first thing to be noted is that when the number of intermediate hops between the MN and the CN increases, the mSCTP latency values continue to increase, while HTM / HIA^{local} latency remains approximately constant. This situation is due to the fact that HTM / HIA^{local} uses the AMU unit to redirect packets to the MN's new location as quick as possible. Then, it updates its association. This approach is completely different from mSCTP that has to update the MN's active association with ADDIP and Set-Primary chunks during the disruption time. Moreover, the HTM^{local} handoff latency remains lower than mSCTP one even if the distance between MN and CN is low. Indeed, with HTM/HIA^{local}, the MN anticipates its address configuration process by using the AMU unit (which is not possible with mSCTP). Recall that the address configuration delay may take over than 500 ms (Mishra et al., 2003)

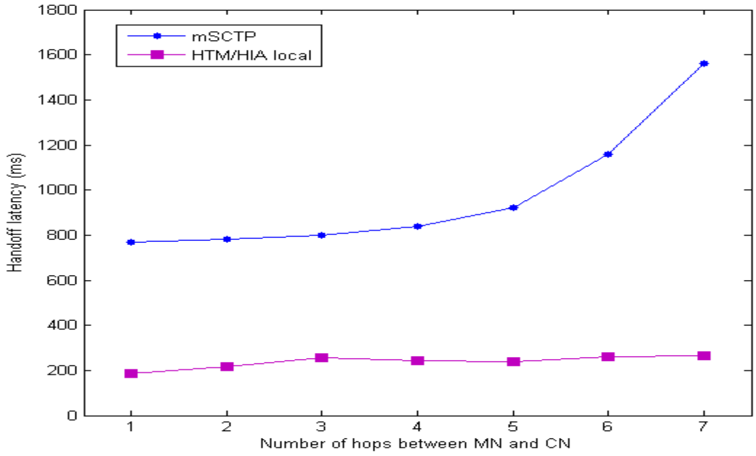


Figure 5.17 Impact of MN-CN distance on handoff latency

Fig. 5.18 shows the throughput pertaining to the time interval (25-40s) following an MN handoff. Note that the *HTM/HIA* throughput is relatively high compared to mSCTP. This is due to the fact that HTM/HIA^{local} uses the IGN/LICS unit to buffer and forward all the traffic to the new MN's location. This traffic obviously includes SACKs which are not lost, unlike what happens with mSCTP. Indeed, the RFC 4960 states that "an endpoint SHOULD transmit reply chunks (e.g., SACK, HEARBEAT ACK, etc.) to the same destination transport address from which it received the DATA or control chunk to which it is replying; and when its pair is multihomed, the SCTP endpoint SHOULD always try to send the SACK to the same destination address from which the last DATA chunk was received". As a result, a number of SACKs transmitted through a previous path fails to reach their destination since the MN has changed its primary IP address. Consequently, unnecessary Congestion Window (CWND) reductions ensue. Under such circumstances, the throughput measured immediately after a handover affected. Accordingly, MN will receive a majority of its SACKs within the RTO time interval (Retransmission TimeOut) since the HTM / HIA^{local} latency is less than 300 ms while the RTO interval is about 1 second.

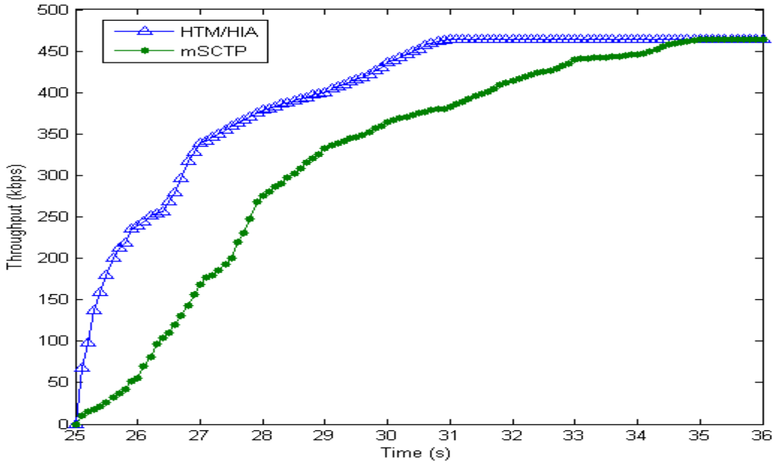


Figure 5.18 Throughput of HTM^{local} vs mSCTP

5.6.3 Numerical results

In this section, we use the previous cost models to illustrate and comment results pertaining to the *HTM/HIA* mobility scheme that runs over the proposed HIA compared to mSCTP.

The list of the parameter values used for our numerical results is shown in Table 5.5.

Table 5.5 Parameters used for performance analysis

Parameters	*Symbols*	Values
Wireless link failure probability	q	0.5
Movement detection delay	T_{MD}	100 ms
L2 handoff delay	T_{L2}	ms50
Address configuration delay	T_{AC}	500 ms
Average queuing delay	ϖ_q	0.1 ms
Wired link bandwidth	B_w	100 Mbps
Wireless link bandwidth	B_{wl}	11 Mbps
Message size	s	296 bytes
Number of AR subnets per AMU/MAP domain	M	4
Average packet arrival per session	A_p	20
Average packets tunneled during a handoff of type (a)	$A_p^{(a)}$	2
Average packets tunneled during a handoff of type (b)	$A_p^{(b)}$	2
Lookup cost at the AMU	l_{AMU}	2
Lookup cost at the LICS	l_{LICS}	2
Packet tunneling cost at the AMU	η_{AMU}	2
Packet tunneling cost at the LICS	η_{LICS}	2
Waiting time before effective data transmission	τ	1 ms

Fig. 5.19 illustrates the total signaling cost as a function of the SMR ratio. When the SMR ratio is inferior to 1, the mobility rate is higher than the session arrival rate; that is why the signaling cost increases for both *HTM/HIA* and mSCTP. This increase becomes more noticeable when the SMR is close to 0. However, the *HTM/HIA* cost remains lower than the mSCTP cost for various values of ε. More specifically, the total signaling cost decreases when ε increases. This means that the introduction of LICS components involves a noticeable diminution of the average signaling cost. On the other hand when the SMR is superior to 1, i,e., the session arrival rate is greater than the mobility rate, the total signaling costs are approximately the same since the association updates are performed less often.

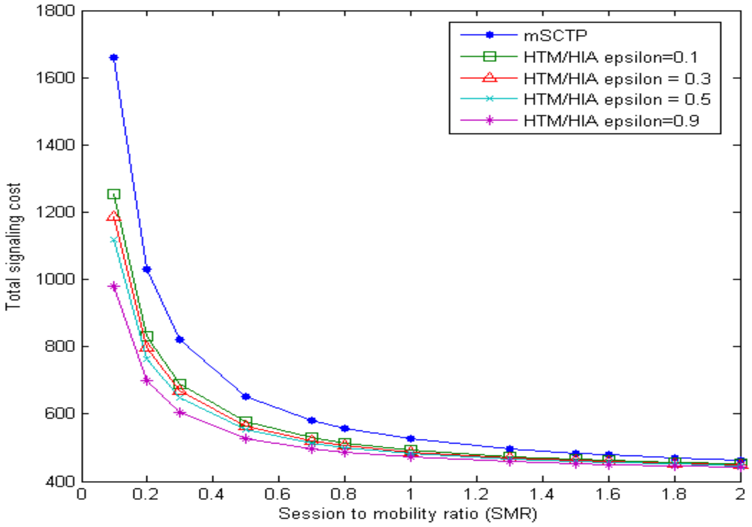


Figure 5.19 Impact of the SMR on the total signaling cost

Fig. 5.20 illustrates the total signaling cost as a function of mobile node velocity. We notice that the estimated signaling cost is proportional to MN's velocity for both *HTM/HIA* and mSCTP. For pedestrian mobility, the total signaling cost is approximately the same for the compared protocols. Nevertheless, the gap between the two signaling costs becomes more and more important depending on the MN's velocity ($v \geq 5ms^{-1}$). This result is to be expected since the MN will perform frequent handoffs when its velocity reaches high values. However, *HTM/HIA* exhibits lower signaling costs since it takes into account local mobility through the IGN/LICS components. This diminution is clearly observed when the probability of performing handoffs through an LICS unit (ε) increases.

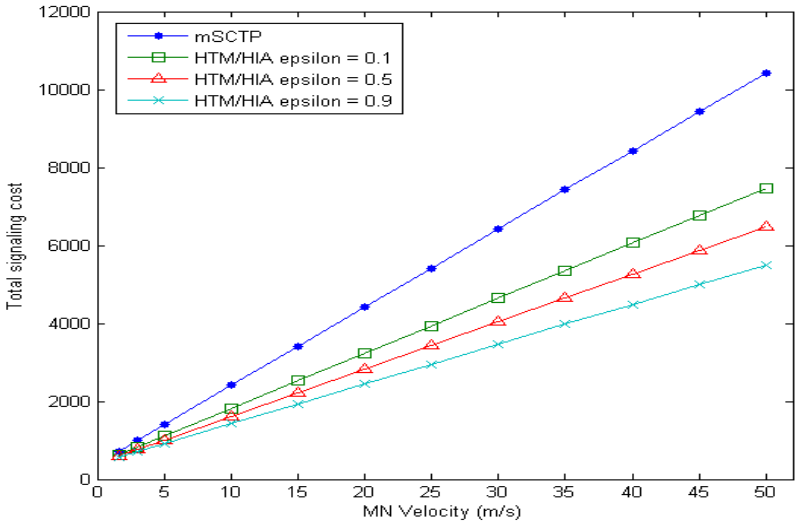


Figure 5.20 Impact of MN velocity on the total signaling cost

Fig. 5.21 shows the behaviour of the total signaling cost as a function of the user density. When this density is low (i.e., *user density* < 0,1), the total signaling cost for both mSCTP and *HTM/HIA* is low and remains approximately the same. However, when this density reaches high values, the AR's border crossing rate increases. Hence, mobile nodes are luckier to perform handoffs. Thus, their corresponding signaling overhead becomes more and more important. Nevertheless, *HTM/HIA* presents lower signaling load amount than mSCTP. This situation is more noticeable for higher values of ε.

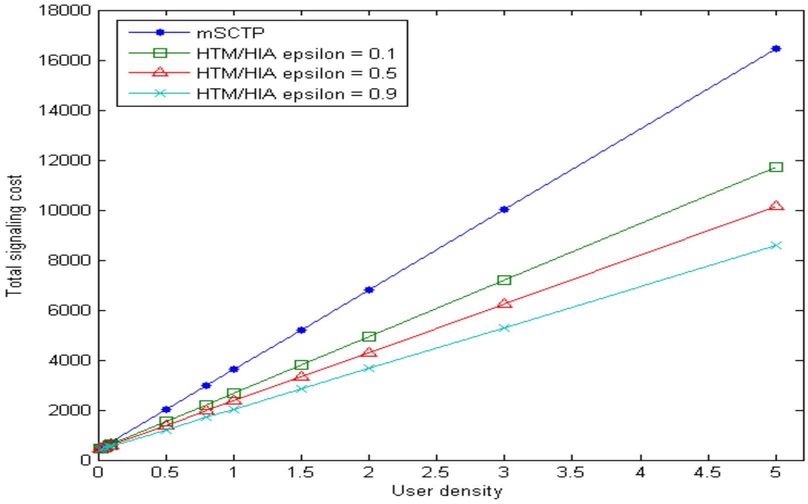


Figure 5.21 Impact of user density on the total signaling cost

Fig. 5.22 shows that the *HTM/HIA* total signaling cost is proportional to the AMU tunneling cost. However, this result remains lower than the mSCTP cost even if with high values for the AMU tunneling cost (i.e., around 20). Recall that all of the processing costs used for our performance analysis are less or equal to 4. On the other hand, mSCTP is not affected by the AMU cost variation since it does not perform traffic redirection. Moreover, we notice that the total signaling cost of *HTM/HIA* decreases when ε increases. This situation is due to the fact that the presence of LICS units limits the amount of signaling messages during handoff periods.

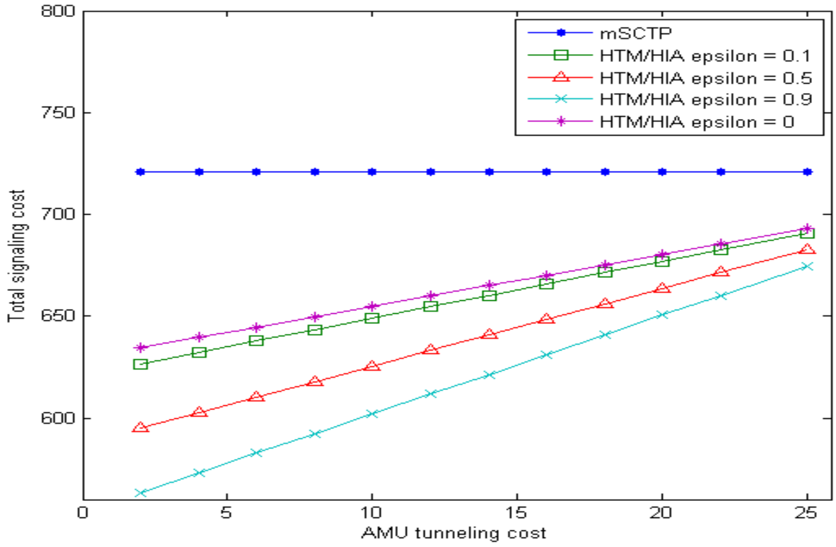


Figure 5.22 Impact of the AMU tunneling cost on the total signaling cost

In Fig. 5.23, we present the average handoff delay as a function of the wireless link delay. We notice that *HTM/HIA* performances are better than mSCTP even though in the absence of LICS-based handoffs (ε =*0*). We also notice that the *HTM/HIA* average latency decreases when the probability of handoffs performed in the presence of an LICS unit increases (i.e., ε=0 to ε=0.9). This means that the introduction of the LICS units is very useful to reduce handoff delays.

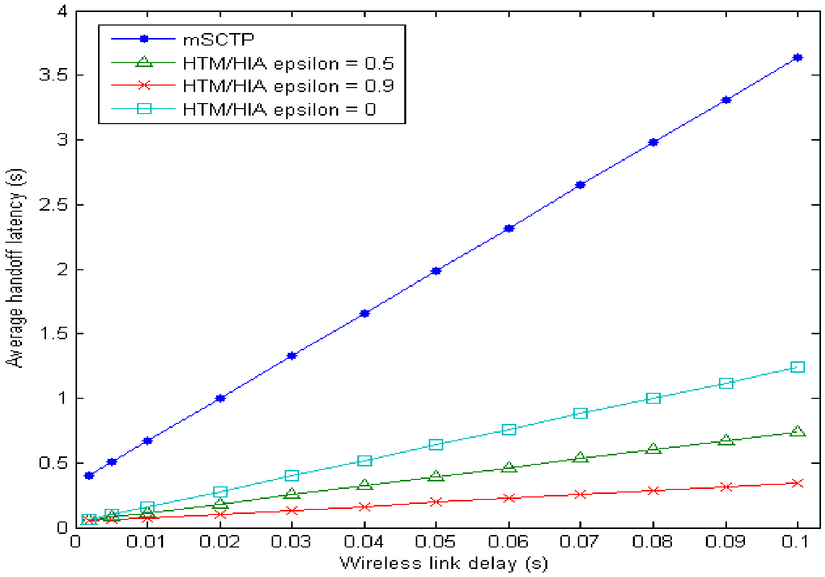


Figure 5.23 Handoff latency as a function of wireless link delay

In Fig. 5.24 we illustrate the impact of performing local handoffs inside an IGN domain (handoff of type (a)) on the average handoff latency. Notice that with *HTM/HIA*, the MN can perform appropriate handoff preparation and can be aware of L2 triggering events. Thus, the MN performs path switching as soon as it is attached to its new location. In this case, handoff latency is limited to L2 link switching (T_{L2}) and association update and packet forwarding (T_{UF}) delays as it is shown in the timing diagram of Fig. 5.12. On the other hand, the latency relevant to mSCTP is high since it does not consider handoff preparation and local handoffs.

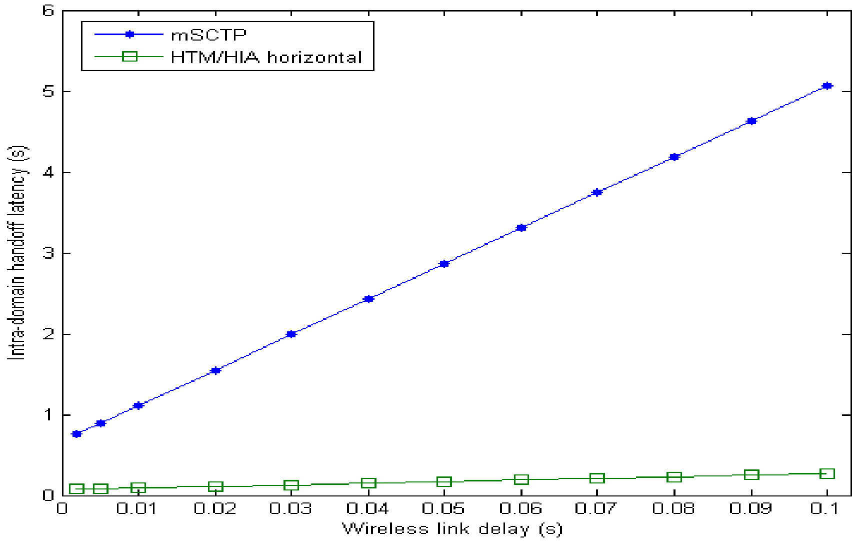


Figure 5.24 *HTM/HIA* and mSCTP intra-system handoffs

Fig 5.25 shows the handoff latency estimated when an MN performs inter-system (vertical) handovers in the presence of an LICS component. It is clear that handoffs completed with LICS present a net improvements of handoff delay compared to the one that uses the standard mSCTP handoff procedure. Thus, the proposed architecture is very useful to alleviate seamless roaming through heterogeneous mobile systems.

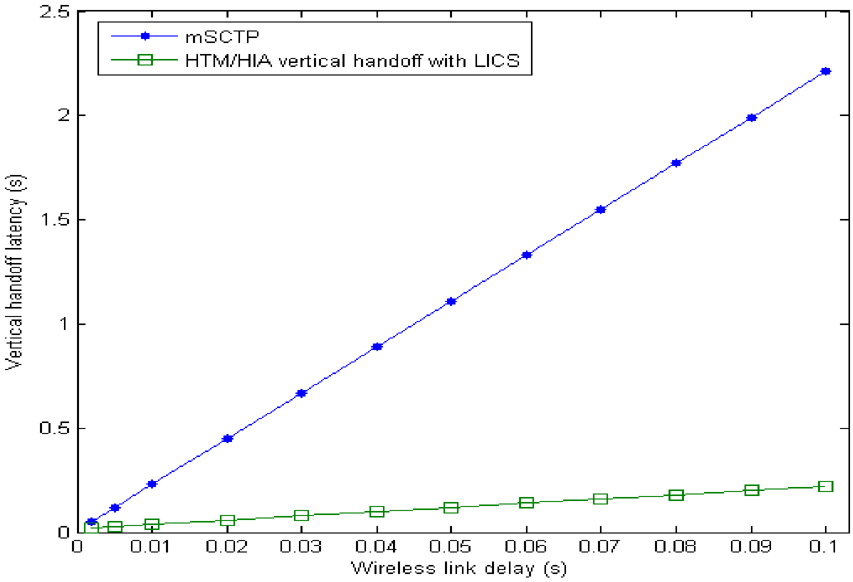


Figure 5.25 LICS based vertical handoffs *vs* mSCTP

Fig. 5.26 shows the behavior of packet loss as a function of packet arrival rate. It is noticed that packet loss remains approximately close to zero for the *HTM/HIA* protocol. However, the mSCTP packet loss increases proportionally to the packet arrival rate. This situation is quite normal since the proposed *HTM/HIA* buffers all packets sent during the handoff period. The observed packet loss for *HTM/HIA* is particularly due to the buffer size. Hence, the proposed HIA contributes to considerably limit packet loss during handoffs performed through homogeneous/heterogeneous networks. On the other hand, all packets exchanged with mSCTP during the handoff period are lost.

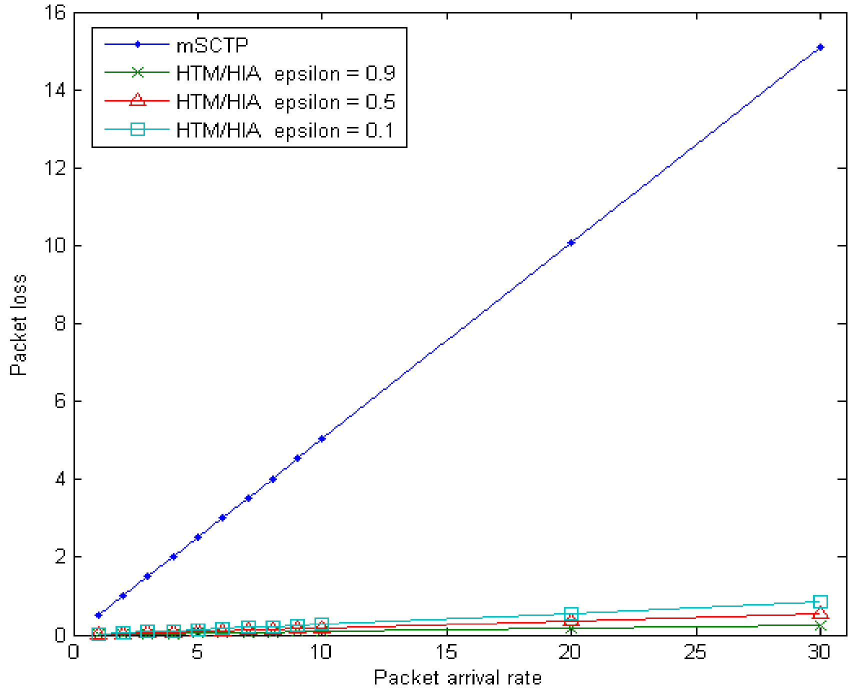


Figure 5.26 Packet loss behavior for different packet arrival rates

As it is illustrated in Fig. 5.27, the handoff blocking probability is more important when the border crossing rate is high. This means that if the MN performs consecutive handoffs in a very short time, it is more likely to have unsuccessful handoffs. However, when the border crossing rate is small, the handoff blocking probability goes down. More specifically, we notice that *HTM/HIA* improves the handoff blocking probability compared to mSCTP. This means that the proposed HIA guarantees a lower handoff blocking likelihood compared to the traditional roaming scenario where the IGN/LICS units are not considered.

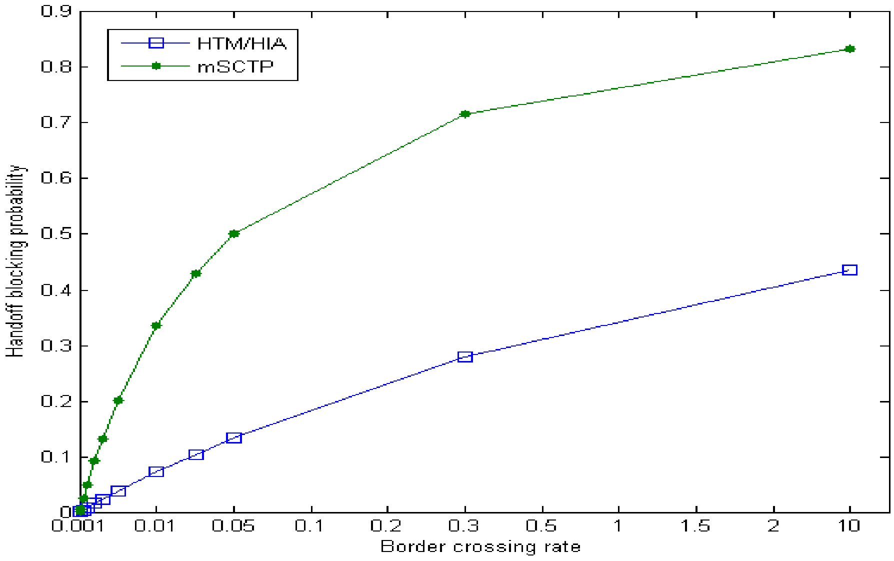


Figure 5.27 Handoff blocking probability *vs* border crossing rate

Fig. 5.28 shows the ratio processing load as a function of the *M* parameter (number of AR subnets in IGN domain). We notice that for different values of *M*, the ratio $\frac{L_{ICS}}{L_{HA}}$ is inferior to 1. This means that the processing load incurred by handoffs at the ICS is lower than the one's at the traditional HA. We also notice that the gap between L_{HA} and L_{ICS} becomes more and more important when the number of subnets increases. We also notice the same behavior when ε increases.

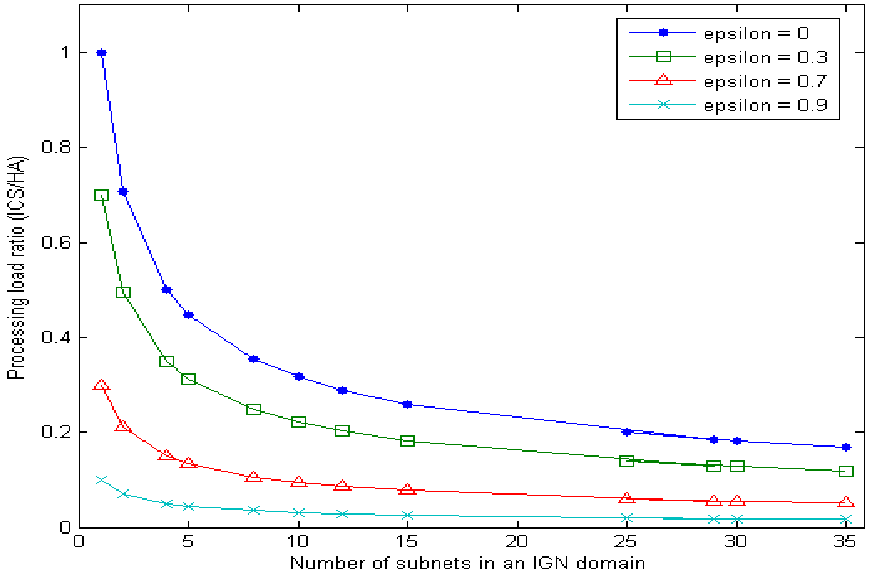


Figure 5.28 Processing load ratio *vs* number of IGN domains

5.7 Conclusion

In this paper, we presented a new hybrid interworking architecture (HIA) which aims to integrate metropolitan mobile networks. HIA introduces an Interworking Cooperation Server (ICS) that operates as an independent authority to ensure billing services and provide context information through heterogeneous technologies. Moreover, the ICS unit reduces considerably the service level agreements (SLAs) since it mediates inter-system authentication rather than using bilateral authentication approach between all the existing networks. The ICS entity manipulates only signaling traffic, so it could handle a large number of operators inside one city. We have also introduced a soft-tight coupling that uses a Local Interworking Cooperation Server (LICS) to render more efficient mobile user roaming and handles adequately real-time traffic. HIA is enhanced with an efficient mobility scheme (*HTM/HIA*) that takes into account context awareness and network selection. In this sense, border nodes (IGN) are enhanced with AMU and

CAS functionalities. Numerical results show that the proposed architecture reduces significantly handoff delays, packet loss, signaling cost and handoff blocking rate. In addition, the processing load at the ICS is considerably lower than the processing load observed at an HA unit. Finally, HIA is scalable and does not require extensive costs for deployment since it is built over existing components.

CHAPITRE 6
DISCUSSION GÉNÉRALE

Ce chapitre se veut une discussion générale relative aux différents points abordés dans cette thèse. Pour ce faire, nous allons, dans un premier temps, présenter une synthèse des travaux réalisés tout en soulignant jusqu'à quel point nos objectifs de recherche ont été atteints. Ensuite, nous mettrons l'accent sur la méthodologie suivie pour mener à bien cette thèse. Enfin, nous effectuerons une analyse des résultats et présenterons les conclusions obtenues de ce travail.

6.1 Synthèse des travaux

En résumé, les travaux de recherche menés dans le cadre de cette thèse ont donné lieu à trois articles de journaux et plusieurs articles de conférences avec comité de lecture. Les articles en question émanent de nos principaux axes de recherche préalablement présentés.

Plus spécifiquement, notre premier objectif de recherche consistait à réaliser une revue de littérature ayant un lien direct avec les problématiques des réseaux mobiles de prochaine génération. Cet objectif a été atteint moyennant une analyse approfondie des différentes approches d'intégration, de gestion de mobilité et de décision de relève qui ont trait aux réseaux 4G. Par ailleurs, nous avons pris le soin de couvrir un large spectre des travaux récemment apparus dans les revues les plus connues de la discipline, et ce, pour prendre connaissance des limitations des travaux existants et s'orienter vers des pistes originales pour notre recherche.

En se basant sur cette revue de littérature, il s'est avéré que la gestion de mobilité demeure un point culminant pour toute éventuelle intégration des systèmes mobiles hétérogènes. Après l'analyse des approches de mobilité proposées au niveau de chacune des couches de la pile de protocole TCP/IP, la couche transport se présente alors comme un choix attractif pour supporter la mobilité dans un environnement 4G.

Dans ce sens, nous avons proposé un mécanisme de gestion de mobilité basé sur le protocole SCTP ainsi que sur sa version mobile communément désignée par mobile SCTP ou mSCTP. Le protocole proposé vise à tirer profit des avantages de la mobilité au niveau transport, en l'occurrence le contrôle de flux et la non dépendance des détails des couches inférieures, tout en adressant le problème de mobilité locale et celui de la dégradation du flux des données échangées durant la phase de relève. Toutefois, dans un environnement 4G, un mécanisme de gestion de mobilité n'aura l'effet escompté que s'il est associé à une stratégie efficace de préparation des relèves. Ceci constitue l'idée de base de notre deuxième contribution. Plus précisément, celle-ci consiste à proposer une stratégie de relève adaptée aux exigences de la 4G. En d'autres termes, ce volet de notre recherche couvre l'analyse de contexte, l'initiation de relève et le choix du meilleur réseau de destination. En effet, les solutions antérieures se contentent de supposer l'existence des informations de contexte sans spécifier ni comment les obtenir ni comment faire face à la réticence des opérateurs mobiles à l'idée de partager leurs données internes. De plus, la puissance du signal reçu (RSS) demeure le paramètre de contexte le plus utilisé pour initier les processus de relèves. En outre, le choix du réseau de destination se restreint à des fonctions de préférence primitives qui ne tiennent compte ni du type de relève ni de la stabilité du réseau de destination. Notre principal objectif avec cette deuxième contribution est de concevoir une nouvelle stratégie de décision de relève qui sera en mesure d'éviter les limitations et les faiblesses susmentionnées.

Enfin, dans le but de mettre à contribution ces deux propositions, nous avons proposé une architecture d'intégration qui se veut évolutive, flexible et facile à déployer. Par ailleurs, cette architecture est conçue de manière à supporter tous les schémas d'intégration conventionnels, en particulier les couplages fort et faible. De plus, pour assurer une itinérance sans coupure tout en respectant la préférence des usagers et les limitations des réseaux visités, cette architecture est couplée avec un mécanisme de mobilité qui se veut bien adapté aux environnements hétérogènes dans la mesure où le schémas de mobilité proposé fait appel aux couches inférieures (L2 & L3) pour assurer une préparation efficace des relèves. En outre, pour offrir un support d'échange interopérable, ladite architecture intègre un mécanisme d'authentification basé sur l'approche de *passeport de sécurité* pour réduire les délais d'authentification. Par ailleurs, cette architecture incorpore un système de facturation visant à garantir le suivi des usagers et permet la mise à jour de leurs profils auprès de leurs opérateurs d'origine. Finalement, la solution introduite permet l'obtention des informations de contexte auprès des systèmes hétérogènes tout en respectant la confidentialité des réseaux intégrés.

6.2 Méthodologie

Une fois l'analyse de littérature terminée, nous avons abordé le premier volet de notre problématique de recherche, à savoir la gestion de mobilité dans un environnement hétérogène. Tout au long de l'élaboration de notre mécanisme de mobilité, nous avons gardé à l'esprit les exigences de la 4G en ce qui concerne l'itinérance globale et la garantie de qualité de service. Dans le but de mettre en exergue la solution proposée, nous avons opté pour une validation par simulation ainsi qu'une modélisation analytique. L'implémentation de notre solution, ainsi que celle avec laquelle nous nous sommes comparés, a été accomplie à l'aide du simulateur ns-2. Toutefois, l'absence de certains modules dans ns-2 nous a obligé à développer de nouveaux modules et de les implémenter pour que nous puissions valider notre proposition dans un environnement de simulation plus proche des conditions réelles. Nous avons également élaboré un modèle analytique dans le but de nous assurer que les résultats empiriques et numériques convergent. De plus, le modèle théorique proposé nous a permis de valider certains aspects que nous n'avons pas pu vérifier à l'aide des simulations.

Afin de bien préparer la phase d'avant relève, nous avons proposé une stratégie de relève basée sur trois points essentiels, à savoir: l'analyse de contexte, l'initiation de relève et le choix du prochain réseau de destination. Pour atteindre ces sous-objectifs, nous avons proposé une architecture répartie pour l'analyse de contexte, une stratégie d'initiation de relève basée sur la logique floue et une nouvelle fonction de préférence adaptée aux environnements multicritères. Quant à la validation de notre proposition, nous avons utilisé une série de tests réalisés en partie à l'aide de "*MATLAB*" ainsi qu'au moyen de modules programmés en C++.

Enfin, nous avons proposé une architecture d'intégration hybride bâtie sur des infrastructures existantes tout en respectant les requis d'une itinérance globale à travers des systèmes sans fil hétérogènes. Afin d'atteindre cet objectif, nous avons introduit l'entité ICS qui opère au niveau du plan de contrôle pour assurer l'authentification, la mise-à-jour des profils de facturation et la supervision des échanges d'informations de contexte. Par ailleurs, nous avons proposé l'utilisation de composants LICS qui représentent des ICS locaux dotés de fonctionnalités de redirection du trafic entre les réseaux intégrés. Finalement, l'architecture conçue supporte aussi bien la mobilité au niveau IP qu'au niveau transport. La validation de l'architecture proposée se base sur un modèle analytique robuste qui tient compte de l'aspect mobile et aléatoire des schémas de mobilité qui peuvent avoir lieu dans un environnement réel.

6.3 Analyse des résultats

Les outils d'analyse des performances utilisés dans le cadre de cette thèse ont permis de vérifier l'efficacité et l'adaptabilité des mécanismes, protocoles et architecture proposés. En effet, il a été démontré, à l'aide de simulations et de modèle analytique, que le mécanisme de mobilité proposé permet une bonne réduction de la moyenne des délais des relèves, de la perte des paquets et de la signalisation sur le réseau. De plus, il a été également prouvé que le débit de données d'après relève est amélioré comparativement aux mécanismes de mobilité utilisant la version standard du protocole mSCTP. Par ailleurs, la stratégie de décision de relève proposée, permet de définir avec plus de précision le type (forcée *vs* volontaire) ainsi que les conditions sous lesquelles une relève sera initiée. De plus, l'utilisation de la logique floue dans cette stratégie favorise la considération de plusieurs paramètres de contexte, et ce, indépendamment du fait qu'ils soient exprimés de façon numérique ou linguistique. Les résultats obtenus, moyennant notre plan de tests, montrent que lorsqu'on se contente uniquement du signal RSS comme critère de base pour initier les relèves, on ignore un nombre important de relèves forcées liées à d'autres critères de contexte. En effet, il a été démontré qu'un nœud mobile peut bien avoir une bonne qualité du signal reçu, mais celle-ci peut être associée à une très faible bande passante, un trafic élevé, un coût monétaire non abordable, etc. Il devient donc clair que la considération seule du critère RSS n'est pas fiable pour déclencher des processus de relève de façon appropriée. En outre le mécanisme d'analyse de contexte proposé, permet de garantir la confidentialité des informations de contexte tout en réduisant la signalisation sur le réseau lors de l'accès aux informations de contexte. De plus, les expériences que nous avons menées concernant le choix du réseau de destination ont montré que la stratégie proposée assure toujours un bon choix de réseau de destination, comparativement aux solutions basées uniquement sur la puissance du signal reçu (RSS).

D'un autre côté, l'architecture introduite se veut bien adaptée aux exigences des réseaux mobiles de prochaine génération. En effet, cette architecture est conçue de manière à réutiliser au maximum les infrastructures existantes, tout en supportant aussi bien la mobilité au niveau réseau qu'au niveau transport. L'analyse des performances effectuée pour valider cette architecture a montré que celle-ci permet de réduire de façon considérable les accords bilatéraux entre les réseaux intégrés. En effet, nous avons mis en place une politique d'accès aux services des réseaux visités basée sur la négociation d'un passeport de sécurité délivré par l'entité ICS qui représente la tierce autorité. De plus, l'itinérance globale des usagers mobiles est devenue plus transparente puisque l'architecture en question assure des relèves sans coupure en garantissant une latence, une perte de paquets et une probabilité de blocage minimale, comparativement aux solutions conventionnelles.

CHAPITRE 7

CONCLUSION

La prochaine génération des réseaux mobiles, communément désignée par 4G, vise à satisfaire les exigences des usagers mobiles en termes d'itinérance sans coupure, de garantie de qualité de service et des préférences des usagers. Pour ce faire, l'intégration et la convergence des systèmes mobiles existants et ceux à venir constituent la base de toute éventuelle coexistence entre technologies hétérogènes. De ce nouveau concept de réseaux mobiles, émane un nombre important de problématiques et de défis qui nécessitent des efforts laborieux pour faire en sorte que le concept des réseaux 4G puisse sortir du cadre théorique à une exploitation réelle. Tout au long de cette thèse, nous avons abordé les éléments de problématique qui ont trait à la mobilité des usagers, à la décision de relève et à l'architecture d'intégration. Dans le présent chapitre, nous mettrons en évidence les principales contributions de cette thèse. Ensuite, nous spécifierons les limitations relatives à nos propositions. Enfin, nous proposerons des recommandations ainsi que les éventuelles extensions de ce travail.

7.1 Récapitulatif des contributions

Le principal objectif de cette thèse était de concevoir et de proposer des solutions de gestion de mobilité et d'intégration des réseaux hétérogènes. Cet objectif a été atteint dans la mesure où cette thèse a donné lieu à plusieurs contributions qui touchent directement les requis majeurs de la 4G.

À titre de récapitulation, les contributions essentielles de la présente thèse se résument comme suit:

- proposition d'un mécanisme de gestion de mobilité de bout en bout qui tient compte de la mobilité locale et globale et qui vise à réduire le délais des relèves, la perte des paquets et charge de signalisation sur le réseau. De plus, le problème de détérioration du flux de données reçues après l'exécution d'une relève a été traité.
- conception d'une architecture d'analyse de contexte qui permet d'assurer la disponibilité des informations à travers des systèmes et des environnements hétérogènes d'une part, et de garantir la confidentialité des informations échangées d'autre part.
- proposition d'une stratégie d'initiation de relève basée sur la logique floue, le but étant de déterminer les conditions opportunes pour initier une relève. De plus, notre solution permet d'identifier le type de relève (forcée *vs* volontaire) à déclencher, ce qui offre une importante marge de manœuvre quant au choix du prochain réseau de destination ou point d'attache.
- développement d'une nouvelle fonction de préférence qui considère un nombre variable de paramètres de contexte et qui tient compte également de la stabilité des réseaux lors du choix d'une destination.
- conception d'une architecture d'intégration interopérable pour les réseaux métropolitains. Par ailleurs, cette architecture est ouverte et peut supporter aussi bien la mobilité au niveau IP qu'au niveau transport.
- proposition d'une version améliorée du protocole HTM de manière à garantir la qualité de service en incluant les phases de préparation des relèves et du choix des réseaux de destination.
- validation des solutions proposées moyennant des simulations et des modèles théoriques.

7.2 Limitations des travaux

La problématique d'intégration des réseaux mobiles demeure ouverte et les solutions proposées jusqu'à date ne font pas l'unanimité dans la communauté scientifique. En conséquence, notre contribution, dans le cadre de cette thèse, ne présume pas être la solution mais elle a pu, à notre avis, apporter des éléments de solutions à certaines des problématiques abordées. Toutefois, notre travail comporte quelques limitations dues à la nature du sujet traité, à la solution proposée et aux plateformes de simulation utilisées.

Une première limitation est due à la nature du sujet abordé où il est difficile, faute de temps, de mener une étude globale sur toutes les problématiques émanant de l'intégration tels que : la mobilité, la sécurité, l'interopérabilité, la garantie de qualité de service, la facturation, l'adaptabilité des terminaux mobiles, etc.

L'introduction de l'unité LICS peut apparaître comme un point de rupture en cas de saturation ou de déni de service. Toutefois, cette crainte demeure présente même si les réseaux intégrés ne passent pas par un LICS. En effet, lors de sa connexion à Internet, n'importe quel réseau passera obligatoirement par une passerelle de sortie. Celle-ci peut donc présenter les mêmes problèmes qu'un LICS. De plus, avec les énormes progrès que connaissent les infrastructures de télécommunication, il est évident que ce genre de composants sera doté de bonnes capacités de calculs et de traitements. En conséquence, les limitations dues au trafic seront peu influente devant la puissance des infrastructures utilisées.

Une autre limitation est liée, cette fois, aux simulateurs utilisés. En effet, ceux-ci n'offrent pas les modules et les fonctionnalités désirés, ce qui nous a obligé à implémenter nos propres modules. Cette façon de faire cible juste notre besoin et n'implémente pas de façon globale les solutions avec lesquelles nous nous sommes comparées. Ceci pourrait avoir un impact sur la validation à grande échelle. De plus, nous aurions bien aimé faire des tests avec des réseaux réels, mais ce genre de validation demeure onéreux et l'accès aux données réelles est loin d'être facile à cause de la réticence des opérateurs à l'idée divulguer leurs données privées.

7.3 Extensions et travaux futurs

Comme nous l'avons préalablement souligné, les problématiques relevant de la prochaine génération des réseaux mobiles demeurent ouvertes et d'actualité. Dans cette section, nous présenterons quelques extensions que nous considérons comme des pistes potentielles pour des travaux futurs pouvant se rapporter directement à la présente thèse.

Une première extension à notre travail sera une comparaison empirique et analytique du mécanisme de mobilité proposé avec ceux des niveaux réseau et applicatif. De plus, il serait extrêmement intéressant d'élaborer un cadre de test, dans lequel nous pourrions étudier des scénarios d'intégration pour identifier les approches de mobilité les plus appropriés pour chacun des scénarios étudiés.

Une autre extension de ce travail serait d'étudier l'impact du mécanisme d'analyse de contexte sur la stratégie de décision de relève que nous avons proposée dans la mesure où notre solution doit être implémentée comme une seule suite de protocole.

Enfin, la proposition d'un mécanisme de sélection des paramètres de contexte les plus appropriés pour une relève serait d'une grande utilité pour optimiser le processus d'itinérance. En d'autres termes, il serait souhaitable d'avoir la possibilité d'identifier à l'avance le nombre de paramètres de contexte à considérer ainsi que la définition de leur priorité vis-à-vis des services engagés par les usagers mobiles.

DIFFUSION DES RÉSULTATS

- **Articles de revues / journaux**

J1)- A. Ezzouhairi, A. Quintero, S. Pierre : "Towards Cross Layer Mobility Support in Metropolitan Networks", *Computer Communication Journal (Elsevier)*, accepté pour publication Septembre 2009.

J2)- A. Ezzouhairi, A. Quintero, S. Pierre : "Adaptive end-to-end Mobility Scheme for Seamless Horizontal and Vertical Handoffs", *Ubiquitous Computing and Communication Journal* , accepté pour publication en Août 2009.

J3)- A. Ezzouhairi, A. Quintero, S. Pierre : "Enhanced Transport layer Mobility Scheme for Seamless Handoffs", soumis à *Computer Communication* (en révision).

J4)- A. Ezzouhairi, A. Quintero, S. Pierre : "Adaptive Decision Making Strategy for Handoff Triggering and Network Selection", *soumis Journal of Computers (JCP Academy Publishers)* Juin 2009.

- **Articles de conférences**

C1)- Ezzouhairi, A. Quintero, S. Pierre : "A new SCTP mobility scheme supporting vertical handover", *Wireless and Mobile Computing, Networking and Communications, 2006 (Wimob'2006)*, IEEE International Conference, Montreal Canada, June 2006, pp 205-211.

C2)- Ezzouhairi, A. Quintero, S. Pierre : "A Fuzzy Decision Making Strategy for Vertical Handoffs", *Canadian Conference on Electrical and Computer Engineering, CCECE 2008*, Niagara Falls Canada, May 2008.

C3)- Ezzouhairi, A. Quintero, S. Pierre : " Towards cross layer based mobility for 4G networks", *Wireless and Mobile Computing, Networking and Communications, 2009 (Wimob'2009)*, IEEE International Conference, Marrakech Morocco, October 2009.

BIBLIOGRAPHIE

3GPP TS, "3GPP System to WLAN Interworking, "System Description (Release 6)," *3GPP TS 23.234 v6.3.0*, Marsh 2004.

3GPP2 TS, "cdma2000-WLAN Interworking, Stage 1 Requirements", *3GPP2 S.R0087-A v1.0*, February 2006.

3GPP2 TS, "cdma2000-WLAN Interworking, Stage 1 Requirements", *3GPP2 S.R0087-A v1.0*, February 2006.

3GPP, 2005) TS 23.228, "IP Multimedia Subsystem (IMS)", stage 2, 2005.

Ahmed, T., Kyamakya, K., Ludwig, M., 2006, "A contexte-aware vertical handover decision algorithm for multimode mobile terminals and its performance". *Proceeding of the IEEE/ACM Euro American Conference on Telematics and Information System (EATIS 2006)*, pp. 19-28.

Atallah, J. G. and Ismail, M., "Future 4G front-ends enabling smooth vertical handovers", *IEEE circuits and Device Magazine*, Vol. 22, No. 1, 2006, pp. 6-15.

AKYILDIZ, I. F., MOHANTY, S., XIE, J., 2005, "A Ubiquitous Mobile Communication Architecture for Next-Generation Heterogeneous Wireless Systems", *IEEE Communication Magazine*, Vol. 43, No. 6, pp. 29-36.

Akyildiz, I. F., McNair, J., Ho, J. S. M., Uzunalioglu, H., Wang, W., 1999, "Mobility Management in Next-Generation Wireless Systems", *Proceeding of the IEEE*, Vol. 87, No. 8, pp. 1347-1384.

Balasubramaniam, S., Indulska, J., 2004, "Vertical handover supporting pervasive computing in future wireless networks", *Computer Communications*, Vol. 27, No. 8, pp. 708-719.

Banergiee, N., Wu, W., Basu, K., Das, S. K., 2004, "Analysis of SIP-based mobility management in 4G wireless networks", *Computer Communications*, Vol. 27, No. 8, pp. 697-707.

Bauman, F. V., Niemegeers, I. G, 1994, "An Evaluation of Location Management Procedures", *Proceeding Of 3rd Annual Int'l Conference Universal Personal Communications (UP' 94)*, September 1994, pp. 359-364.

Buddhikot, M., Chandranmenon, G., Han, S., Lee, Y. W., Miller, S., Salgarelli, L., 2003, "Integration of 802.11 and third-generation wireless data networks", *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE*, Vol. 1, No. 1, pp. 503-512.

Campbell, A.T., Gomez, J., Valko, A.G., "An overview of cellular IP", *Wireless Communications and Networking Conference*, WCNC 1999 IEEE, Vol. 2, pp. 606-610.

Calvagna, A., Di Modica, G., 2004, "A user-centric analysis of vertical handovers", *Proceedings of the second ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, pp. 137-146.

Cavalcanti, D., Agrawal, D., Cordeiro, C., Xie, B., Kumar, A., 2005, "Issues in Integrating Cellular Networks, WLAN, and MANETs : A Futuristic Heterogeneous Wireless Network", *IEEE Wireless Communication*, Vol. 12, No. 3, pp. 30-41.

Chen, W., Liu, J., Huang, H., 2004, "An Adaptive Scheme for Vertical Handoff in Wireless Overlay Networks", *Proceedings on the 10th International Conference on Parallel and Distributed Systems*, pp. 541-548.

Chi-Hsing, H., Bernard, J. C., 1999, "Fuzzy multiple attribute decision making using a simplified centroid-based arithmetic process", *International Journal of Industrial Engineering : Theory Applications and Practice*, Vol. 6, No. 1, pp. 61-71.

Droms, R., 1997, "Dynamic Host Configuration Protocol", *IETF RFC 2131*, March 1997.

Fang, Y., 2003, "Movement-Based Mobility Management and Trade Off Analysis for Wireless Mobile Networks", *IEEE Transactions on Computers*, Vol. 52, No. 06, pp. 791-803.

Fu, S., Atiquzzaman, M., "Hierarchical location management for transport layer mobility", *Tech. Rep. TR-OU-TNRL-05-105*, University of Oklahoma, Telecommunication & Network research Lab, www.cs.ou.edu/~netlab, Feb 2005.

Fracchia, R., Casetti, C., Chiasserini, C. F., Meo, M., 2007, "WiSE: Best-Path Selection in Wireless Multihoming Environments", *IEEE Transactions on Mobile Computing*, Vol. 6, No. 10, pp. 1130 – 1141.

Frattasi, S., Fathi, H., Fitzek, F. H. P., Prasad, R., Katz, M. D., 2006, "Defining 4G technology from the users perspective", *IEEE Network*, Vol. 20, No. 1, pp. 35-41.

Fu, F. S., Atiquzzaman, M., Ma, L., Ivancic, W., Lee, Y., Jones, J. S., Lu, S., 2004, " TraSH : A Transport Layer Seamless Handover for Mobile Networks", *Technical Report: OU-TNRI-04-100*.

Fu, F. S., Ma, L., Atiquzzaman, M., Lee, Y., "Architecture and performance of SIGMA: A seamless mobility architecture for data networks", *40th IEEE International Conference on Communication (ICC)*, Seoul, Korea, May 2005, pp. 3249-3253.

Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B., "Proxy Mobile IP", *RFC 5213*, August 2008.

Guo, Q., Zhu, J., Xu, X., 2005, "An adaptive multi-criteria vertical handoff decision algorithm for radio heterogeneous network", *IEEE International Conference on Communications (ICC 2005)*, Vol. 4, pp. 2769–2773.

Gustafsson, E., Jonsson, A., 2003, "Always Best Coonnected", *IEEE Wireless Communications*, Vol. 10, No. 1, pp. 49-55.

Handley, M., Schulzrinne, H., Schooler, E., Rosenberg, J., 1999, "SIP: Session Initiation Protocol", *IETF RFC 2543.*

Havinga, P. J. M., Smit, G. J. M., Wu, G., Vogniild, L. K., 2001, "The SMART Project: Exploiting the Heterogeneous Mobile World", *International Conference on Internet Computing*, Las Vegas USA, June 2001.

Hsieh, H., Kim, K., Zhu, Y., Sivakumar, R., 2003, "A receiver-centric transport protocol for mobile hosts with heterogeneous wireless interfaces", *ACM MobiCom, San Diego*, pp. 1-15.

Huitema, C., "Multi-homed TCP ", draft-huitema-multi-homed-0.txt, 1995.

Hwang, C., Yoon, K., 1981, "Multiple Attribute Decision Making", Springer-Verlag.

Inter-PLMN backbone guidelines, *GSM association classifications*, version 3.4.0, March 2003.

Iyengar, J. R., Amer, P. D., Stewart, R., 2006, "Concurrent Multipath Transfer Using SCTP Multihoming Over Independent End-to-End Paths", *IEEE Transactions on Mobile Computing*, Vol. 14, No. 5, pp. 951-964.

Jha S., Mukherjee, A., 2004, "Advances in future mobile/wireless networks and services ", *Computer Communications*, Vol. 27, No. 8, pp. 695-696.

Johnson, D., Perkins, C., Arkko, J., 2004, "Mobility Support in IPv6", *IETF RFC 3775.*

Jung, H. Y., Kim, E. A., Yi, J. W., Lee, H. H., 2005, "A scheme for supporting fast handover in hierarchical mobile IPv6 networks", *ETRI Journal*, Vol. 27, No. 6, pp. 798–801.

Kassar, M., Kervella, B., Pujolle, G., 2008, "An overview of vertical handover decision strategies in heterogeneous wireless networks", *Computer Communications*, Vol. 31, No. 10, pp. 2607-2620.

Koodli, G., 2005, "Fast handovers for mobile IPv6", *IETF RFC 4068.*

Koh, S. J., Chang, M. J., Lee, M., 2004, "mSCTP for soft handover in transport layer", *IEEE Communication Letters*, Vol. 8, No. 3, pp. 189-191.

Lassoued, I., Bonnin, J., Ben Hamouda, Z., Belghith, A., 2008, "A methodology for evaluating vertical handoff decision mechanisms", IEEE 7th International Conference on Networking, pp. 377-383.

Leibsh, M., Singh, A., Chaskar, H., Funato, D., Shim, E., 2005, "Candidate Access Router Discovery (CARD)", *IETF RFC 4066*, July 2005.

Ma, L., Yu, F., Leung, V. C. M., Randhawa, T., 2004, "A new method to support UMTS/WLAN vertical handover using SCTP ", *IEEE Wireless Communications*, Vol. 11, No. 4, pp. 44-51.

Makaya, C., Pierre, S., 2007, "An Interworking Architecture for Heterogeneous IP Wireless Network", ICWMC 2007*, Third international conference on wireless and mobile communications*, 16-21.

Makela, J., Ylianttila, M., Pahlavan, K., 2000, "Handoff decision in multiservice networks*", 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, pp. 655–659.

Maltz, D., Bhagwat, P., 1998, "MSOCKS: An architecture for transport layer mobility", *INFOCOM San Francisco*, pp. 1037-1045.

McNair, J., Akyildiz, I. F., Bender, M. D., 2001, "Handoffs for Real-time Traffic in Mobile IP version 6 Networks", *Proceedings of IEEE GLOBECOM*, Vol. 6, pp. 3463-3467.

McNair, J., Zhu, F., 2004, "Vertical Handoffs in Fourth-Generation Multinetwork Environment", *IEEE Wireless Communications*, Vol. 11, No. 3, pp. 8-15.

Mishra, A., Shin, M., Arbaugh, W., 2003, "An empirical analysis of the IEEE 802.11 MAC layer handoff process", *ACM SIGCOMM Computer Communication Review*, Vol. 33, No. 2, pp. 93-102.

Nasser, N., Hasswa, A., Hassanein, H., 2006, "Handoffs in Fourth Generation Heterogeneous Networks", *IEEE Communication Magazine*, Vol. 44, No. 10, pp. 96-103.

Ormond, O., Murphy, J., Muntean, G., 2006, "Utility-based intelligent network selection in beyond 3G systems", *IEEE International Conference on Communications (ICC 2006)*, vol. 4, pp. 1831– 1836.

Pahlavan, K., Krishnamurthy, P., Hatami, A., Ylianttila, M., Makela, J. P., Pichna, R., Vallstron, J., 2000, "Handoff in hybrid mobile data networks", *IEEE wireless communications*, Vol. 7, No. 2, pp. 34-47.

Perkins, C., 2002, "IP Mobility Support for IPv4". *RFC 3344*.

Pe´rez-Costa, X., Torrent-Moreno, M., Hartenstein, H., 2003, "A performance comparison of mobile IPv6, hierarchical mobile IPv6, fast handovers for Mobile IPv6 and their combination", *ACM Mobile Comp. and Comm. Rev*, Vol. 7, No 4.

Pierre, S., 2007, "Réseaux et Systèmes Informatiques Mobiles: Fondements, Architectures et Applications". *Presses Internationales Polytechnique*, Montréal, Édition revue et augmentée.

Qiang, G., Jie, Z., Xu, X., 2005, "An adaptive multi-criteria vertical handoff decision algorithm for radio heterogeneous network", *ICC IEEE International Conference on Communications*, Vol. 4, pp. 2769 – 2773.

Quiqyang, S., Jamalipour, A., 2005, "A network selection mechanism for next generation networks", *International Conference of Communications (ICC 2005)*, Vol. 2, pp. 1418-1422.

Ramjee, R., Varadhan, K., Salgarelli, L., Thuel, S. R., Wang, S. Y., La Porta, T., “HAWAII: A domain-based approach for supporting mobility in widearea wireless networks”, *IEEE/ACM Transactions on Networking*, Vol. 10, No. 3, June 2002, pp. 396-410.

Saaty, T., 1990, "How to Make a Decision: the Analytic Hierarchy Process", *European Journal of Operational Research*, Vol. 48, No. 1, pp. 9-26.

Scharf, M., Kiesel, S., 2006, "NXG03-5: Head-of-line Blocking in TCP and SCTP: Analysis and Measurements", *Global Telecommunications Conference GLOBECOM '06 IEEE*, pp. 1-5.

Shaojian, F., Atiquzzaman, M., 2004, "SCTP: state of the art in research, products, and technical challenges", *Communications Magazine, IEEE*, Vol. 42, No. 4, April 2004, pp. 64-76.

Siddiqui, F., Zeadally, S., 2006, "Mobility management across hybrid wireless networks: Trends and challenges", *Computer Communications*, Vol. 29, No. 9, pp. 1363-1385.

Snoeren, A., Balakrishnan, H., 2000, "An end-to-end approach to host mobility", *ACM MobiCom Boston*, pp. 155-166.

Soliman, H., Castelluccia, C., El-Malki, K., Bellier, L., 2005, "Hierarchical mobile IPv6 mobility management (HMIPv6)", *IETF RFC 4140.*

Stewart, R., 2007, "Stream Control Transmission Protocol", *IETF RFC 4960.*

Stewart, R., Xie, Q., Tuexen, M., Maruyama, S., Kozuka, M., 2007, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration", *IETF RFC 5061.*

Tansu, F., Salamah, M., 2006, "On the vertical handoff decision for wireless overlay networks", *Computer Networks, 2006 International Symposium on*, pp. 111 – 115.

Thomson, S., Narten, T., 1998, "IPv6 Stateless Address Autoconfiguraton", IETF RFC 2462.

Wang, H., Katz, R., Giese, J., 1999, "Policy-Enabled Handoffs Across Heterogeneous Wireless Networks", *Second IEEE Workshop on Mobile Computing system and Applications*, pp. 51-60.

Wang, W., Akyildiz, I. F., 2000, "Intersystem Location Update and Paging Schemes for Multitier Wireless Networks", *Proceeding of ACM MOBICOM*, pp. 99-109.

Wang, W., Akyildiz, I.F., 2001, "A New Signaling Protocol for Intersystem Roaming in Next-Generation Wireless Systems", *IEEE Journal on Selected Area in Communication (JSAC)*, Vol. 19, No. 10, pp. 2040-2052.

Wei, Q., Farkas, K., Prehofer, C., Mendes, P., Plattner, B., 2006, "Context-Aware Handover Using Active Network Technology", *Computer Networks*, Vol. 50, No. 15, pp. 2855-2872.

Wei, S. H., Huang, M. H., Long, W., 2005, "Research on the scheme supporting mobility of TCP application", *Journal of Chongqing University of Posts and Telecommunication*, Vol. 17, No. 1, pp. 117-120.

Xiao, Y., Pan, Y., Li, J., 2004, "Design and Analysis of location Management for 3G Cellular Networks", *IEEE Transactions on parallel and distributed systems*, Vol. 15, No. 04, pp. 339-349.

Young-Jou, L., Ting-Yun, L., Ching-Lai, H., 1994, "TOPSIS for MODM", *European Journal of Operational Research*, Vol. 76, No. 3, pp. 486-500.

Zadeh, L., 1972, "A fuzzy set theoretic interpretation of linguistic hedges", *J. Cybernetics*, Vol. 2, No. 2, pp. 4-34.

Zeadally S., Siddiqui F., 2007, "An Empirical Analysis of Handoff Performance for SIP, Mobile IP and SCTP protocols", *Wireless Personal Communications*, pp. 589-603.

Zhang, J., Chan, H., Leung V., 2007, "A SIP-based Seamless-Handoff (S-SIP) Scheme for Heterogeneous Mobile Networks", *Proceedings of the IEEE Wireless Communications and Networking Conference*, Hong Kong, pp. 3946-3950.

Zhang, W., Jaehner, J., Dolzer, K., 2003, "Design and Evaluation of a Handover Decision Strategy for 4th Generation Mobile Networks", *Vehicular Technology Conference*, 2003. VTC 2003-Spring. The 57th IEEE Semi annual, Vol. 3, No. 3, April 2003, pp. 1969-1973.