Gestion de la mobilité et allocation de ressources dans les réseaux multiservices sans fil

Rola Naja

To cite this version:

Rola Naja. Gestion de la mobilité et allocation de ressources dans les réseaux multiservices sans fil. Réseaux et télécommunications [cs.NI]. Télécom ParisTech, 2003. Français. NNT : . tel-00005726

HAL Id: tel-00005726
https://pastel.hal.science/tel-00005726

Submitted on 5 Apr 2004

Thèse

présentée pour obtenir le grade de docteur
de l'École Nationale Supérieure
des Télécommunications

Spécialité : Informatique et Réseaux

Rola Naja

Gestion de la mobilité et allocation de ressources dans les réseaux multiservices sans fil

Soutenue le 22 septembre 2003 devant le jury composé de:

Guy Pujolle	Président
Pascal Lorenz	Rapporteurs
Stéphane Ubéda	
Isabelle Demeure	Examinateurs
Heba Koraitim	
Paul J. Kuehn	
Samir Tohmé	Directeur de Thèse

Remerciements

A l'aboutissement d'une thèse, il est d'usage de remercier en premier lieu l'encadrant de thèse. Mes premiers hommages vont à mon professeur Samir Tohmé, non pas pour me plier aux coutumes mais pour exprimer ma profonde gratitude envers lui. J'aimerais lui addresser mes plus vifs remerciements pour la liberté qu'il m'a accordée, pour la confiance qu'il avait en moi et pour m'avoir patiemment écoutée surtout lorsque j'ai commencé à avancer dans le long chemin de la recherche avec des pas incertains et hésitants...Je tiens à lui dire que j'ai beaucoup appris de lui tant sur le plan "relationnel" que sur le plan scientifique...

Son calme à toute épreuve et sa juste appréciation des évènements ont su m'apaiser dans mes périodes de doute les plus dures...Il a contribué à me donner le goût de la recherche et je lui suis reconnaissante de m'avoir, en quelque sorte, mis "le pied à l'étrier"...

J'aimerais par ailleurs exprimer ma gratitude à Mr. Michel Riguidel, directeur du département INFRES, pour avoir permis de mener à bout ce travail de recherche dans les meilleures conditions.

Je tiens également à remercier les membres de mon jury de thèse. Je suis reconnaissante envers Professeur Pascal Lorenz et Professeur Stéphane Ubéda pour avoir bien voulu rapporter sur ma thèse. Mes remerciemens vont également aux examinateurs de ma thèse qui ont eu l'amabilité d'examiner ma thèse. Je pense, plus particulièrement au Professeur Paul Kuehn, au Professeur Guy Pujolle, à Dr. Isabelle Demeure et à Dr. Heba Koratitim. Je n'oublierai pas les conseils précieux de Heba qui a su donner à ce mémoire son poli final...

Bien d'autres noms mériteraient d'être cités, notamment les membres du département INFRES. Je pense particulièrement à Dr. Houda Labiod et à Dr. Nicolas Puech pour m'avoir aidée au tout début de ma thèse.

Mes remerciements seraient sans doute incomplets si je ne cite pas:

Mes parents sans qui ce travail n'aurait jamais vu le jour...Je leur suis infiniment reconnaissante pour leur amour, leur soutien moral et leurs encouragements à être toujours la meilleure. Qu'ils trouvent dans ce mémoire le fruit de leur travail!

Ma mère de coeur Fackher. Bien au delà des distances qui nous séparent, j'aimerais pouvoir te dire que ta présence me manque plus que jamais. Tu n'es plus de ce monde, mais je vois toujours ton beau sourire et tes yeux si doux...A toi, je dédie ce mémoire.

Mes soeurs Hala, Dania, Ghinwa et Maya qui ont parsemé ma vie de milliers de roses... Je te remercie Dania pour avoir toujours été là pour moi surtout dans mes débuts. Tu m'as toujours écoutée avec le grand coeur d'une soeur, l'oreille attentive d'une confidente et la tendresse attentionnée d'une mère... Grâce à ton soutien moral salvateur dans certains moments difficiles, tu m'as fait oublier que le Liban est si loin!
Un remerciement particuler est dédié à Hala pour m'avoir vivement encouragée à faire une thèse et pour m'avoir souvent gâtée avec ses petites attentions affectueuses... Merci Nouna pour ton humour inébranlable, pour la générosité de ton coeur et pour tes promenades de “choupette”... Merci Mimosa pour tes chamailleries affectueuses, tes histoires originales et tes qualités de coeur. La balle est dans ton camp pour rentrer dans le club des cinq docteurs...Je te passe le flambeau...A toi maintenant!

Une attention toute spéciale s'addresse à Driss et Hossam pour leur soutien moral et leurs encouragements.

Je garde dans ma mémoire des souvenirs mémorables et précieux de mes deux anges gardien, Nour et Chady. Je leur dédie des voeux aussi purs et aussi sincères que leurs rires cristallins et leurs mignons babillages...

Une pensée particulière est dédiée aux étudiants du laboratoire INFRES qui grâce à leur humour, et à leur joie de vivre ont ponctué ces années de thèse d'instants mémorables.

Je pense particulièrement à Ouahiba qui a vécu avec moi les divers moments tristes et joyeux de ma vie de doctorante...Je suis heureuse d'avoir côtoyé Stuart; Sa bonne humeur, son original humour australien, sa sympathie et ses charmantes histoires, parfois un peu “voulgaires” -), ne lui ont jamais fait défaut et ont souvent égayé mes pauses café...Je suis ravie de porter dans mon coeur la solide amitié de Paul qui m'a apporté beaucoup de joie...J'ai apprécié la gentillesse de Hazar, mon collègue de bureau.

Finalement, je tiens à rendre un hommage particulier à Gisela Melo, ma prof de guitare. Sa musique latino-américaine et le rythme nostalgique et chaleureux de ses mélodies ont ajouté une note originale et agréable à mes heures de travail. Ses cours m'ont souvent projetée dans un monde si différent de celui des réseaux et des files d'attente...

Ne dites pas: "J'ai trouvé la vérité", mais plutôt: "J'ai trouvé une vérité."
Ne dites pas: "J'ai trouvé le chemin de l'âme",
mais plutôt: "J'ai croisé l'âme qui marchait sur mon chemin."...

Khalil Gibran, "Le Prophète"

A mon Rayon de Soleil

Résumé

Introduction

Durant les dernières années, nous avons été témoins d'une croissance spectaculaire dans l'intérêt mondial pour les communications mobiles. La demande croissante en services et le développement de la technologie sans fil continueront leur progression dans les années à venir.

Avec l'augmentation du nombre d'utilisateurs mobiles et l'évolution rapide des réseaux mobiles sans fil, les demandes des utilisateurs en terme de qualité de service (QoS) deviennent de plus en plus exigeantes. Ceci nécessite l'introduction des méthodes avancées de gestion de la ressource radio, d'autant plus que l'interface radio représente le goulet d'étranglement des réseaux mobiles. Dans cette optique, un protocole d'allocation de ressources doit pouvoir gérer efficacement la bande passante tout en fournissant la qualité de service à différentes classes de service.
D'autre part, la gestion de la mobilité constitue un important défi technique à relever. En effet, un protocole de mobilité efficace doit pouvoir empêcher la terminaison forcée de l'appel et permettre l'exécution des applications d'une manière transparente au mouvement de l'utilisateur.

Nos travaux réalisés portent sur la gestion des ressources radio et la mobilité dans les réseaux multiservices sans fil. Les environnements qui sont visés dans ce travail concernent les réseaux de seconde, troisième et quatrième génération.

Dans ce résumé, nous essayerons de mettre l'accent sur les principes de base de nos travaux de recherche et les principaux résultats et conclusions obtenus durant la thèse.

Principes de base de la thèse

Ce travail de thèse se focalise sur la fourniture de qualité de service et la gestion de la mobilité dans les réseaux multiservices sans fil. Avant d'identifier les principales contributions de la thèse, nous présentons les principes de base de notre travail:

- L'utilisation de la ressource radio doit être maximale. En effet, la bande passante radio représente un véritable goulet d'étranglement et par conséquent devrait être bien dimensionnée.

- Dans les réseaux sans fil, le contrôle d'admission traite les nouveaux appels et les appels handover. Puisque la terminaison forcée de l'appel, dûe principalement à l'échec de handover, est mal perçue par les utilisateurs, les appels handover requièrent une étude particulière. Dans nos travaux réalisés, nous nous efforçons de minimiser la probabilité de terminaison forcée des appels.

- La fluctuation dans la disponibilité des ressources est plus importante dans les réseaux sans fil que dans les réseaux filaires. Ceci est principalement dû à la mobilité. Dans ce travail de thèse, nous dimensionnons les réseaux sans fil. Nous ne traitons pas les problèmes du réseau coeur et nous nous penchons sur les problématiques des réseaux d'accès.

- Afin de palier aux problèmes de fluctuation des ressources, nous préconisons l'approche d'adaptabilité dans les réseaux sans fil. Avec cette approche, les applications multimédias devraient être adaptables et capables de prendre en compte l'état du réseau. En d'autres termes, grâce à l'adaptabilité la QoS de bout en bout est élaborée conjointement entre l'application et le réseau afin de satisfaire l'utilisateur de manière à respecter son contrat de service.

  Cette notion d'adaptabilité est présente avec l'apparition des terminaux intelligents. Ces terminaux sont conçus avec des technologies avancées leur permettant de réagir avec le réseau.

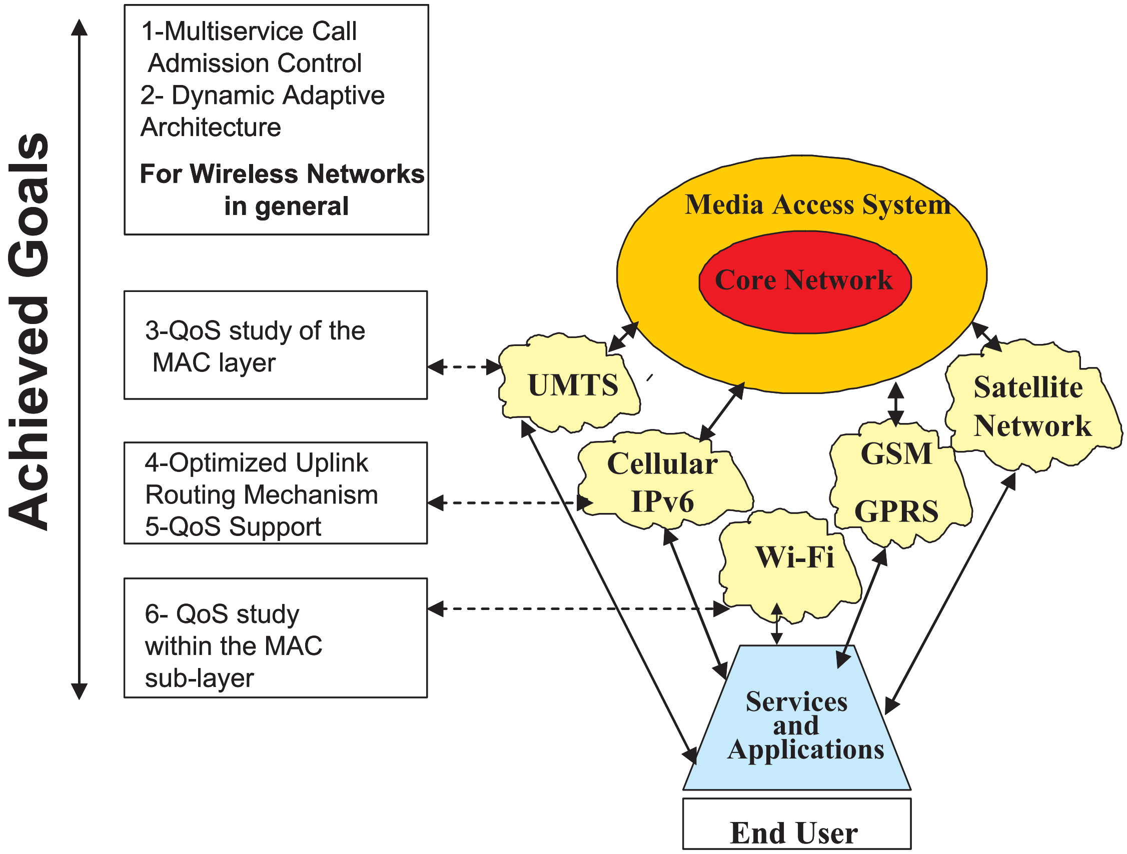

Achieved Goals
1-Multiservice Call Admission Control
2- Dynamic Adaptive Architecture
For Wireless Networks in general
3-QoS study of the MAC layer
4-Optimized Uplink Routing Mechanism
5-QoS Support
6- QoS study within the MAC sub-layer
Media Access System
Core Network
UMTS
Satellite Network
Cellular IPv6
GSM
GPRS
Wi-Fi
Services and Applications
End User

Figure 1: Les contributions de la thèse

Contributions

La gestion de la ressource radio est de loin le problème critique des réseaux cellulaires. De très nombreuses études ont déjà été menées sur ce problème. Cependant, il est loin d'être résolu convenablement dans toutes ses formes.

L'évolution actuelle des réseaux de la troisième génération rend cette problématique encore plus complexe en incluant la notion de multiservices et en envisageant de mixer les technologies. En effet, la prochaine génération des réseaux mobiles sera basée sur des réseaux d'accès sans fil hétérogènes connectés à un réseau coeur tout IP. Les réseaux d'accès WLAN, GSM, GPRS, UMTS et satellitaires feront partie de la classe des réseaux d'accès sans fil de cette future génération de réseaux mobiles.
Ces réseaux futurs offriront différents types de services et différents types de terminaux aux utilisateurs mobiles qui auront différents profils de mobilité et de qualité de service.

Dans cet esprit, nous avons mené des études portant sur la qualité de service et la gestion de mobilité dans certains réseaux d'accès de la prochaine génération des réseaux mobiles. Nos travaux de recherche participent donc à un effort de compréhension dans un domaine aussi vaste que varié. La mobilité verticale, qui gère l'itinérance des utilisateurs se déplaçant d'un réseau d'accès à un autre, fera partie des pistes intéressantes à explorer dans nos travaux futurs.

Nos travaux de recherche apportent six contributions (figure 1). Les deux premières concernent les réseaux multiservices sans fil en général. La troisième et la quatrième contribution traitent les réseaux IP sans fil. La cinquième vise les réseaux IEEE 802.11. Quant à la sixième, elle traite la qualité de service dans les réseaux UMTS.

Plus précisément,

- La première contribution concerne l'étude d'un contrôle d'admission multiservice traitant quatre classes de service dans les réseaux mobiles sans fil.
- La seconde contribution introduit une architecture dynamiquement adaptable DYNAA dont le but est d'absorber les fluctuations du réseau en terme de capacité en introduisant de la dynamique dans la gestion de ces ressources.
- Puisque les réseaux futurs seront probablement basés sur la technologie IP, nous proposons d'étudier le protocole Cellular IPv6. Nous menons ainsi une étude d'optimisation du routage dans le sens montant pour deux mobiles du même réseau Cellular IPv6. Nous traitons également le problème de handover dans ce type de réseaux. Ceci constitue l'axe de recherche de la troisième contribution.
- La quatrième contribution vise également les réseaux Cellular IPv6 et traite le support de qualité de service dans ce type de réseaux.
- La cinquième contribution concerne l'étude de la qualité de service au niveau de la couche Medium Access Control (MAC) des réseaux IEEE 802.11.
- La sixième contribution est une étude de la qualité de service au niveau de la couche MAC des réseaux UMTS.

Dans ce qui suit, nous allons présenter la problématique et le contexte de chacune des contributions citées ci-dessus, en mettant l'accent sur les résultats obtenus.

Etude d'un contrôle d'admission multiservice et gestion de la mobilité dans des réseaux sans fil

Contexte

Le contrôle d'admission (CAC) est un élément clé pour la fourniture de la qualité de service dans les réseaux mobiles sans fil. Un CAC efficace devrait gérer la ressource rare tout en fournissant les différents services ayant différents critères de qualité de service. Ces derniers impliquent l'existence de différents niveaux de priorité.

Ainsi, les appels entrants dans une cellule ou nouveaux appels devraient avoir moins de priorité que les appels handover. Parallèlement, les requêtes temps réel requièrent des exigences assez strictes en terme de délai et de gigue; elles sont donc prioritaires aux requêtes non-temps réel.

Dans cette optique là, l'un des importants enjeux dans la conception d'un contrôle d'admission dans les réseaux multiservices sans fil est la gestion de la ressource radio en respectant les différents niveaux de priorité des appels servis.

Proposition et Résultats

Nous proposons un CAC servant quatre classes de service. Dans cette étude, nous séparons les nouveaux appels par rapport aux requêtes handover et les demandes concernant les services voix par rapport aux services de donnée. Ainsi nous considérons quatre classes de service: les nouveaux appels de donnée (NC-data), les nouveaux appels voix (NC-voice), les requêtes handover de donnée (HO-data) et les requêtes handover voix (HO-voice).

L'algorithme CAC étudié propose un système de priorité pour les quatre classes de service dans le sens croissant suivant: nouvel appel de donnée (NC-data)/ nouvel appel voix (NC-voice)/ requête handover de donnée (HO-data)/ requête handover voix (HO-voice).

Afin d'établir la priorité entre les nouveaux appels voix sur les nouveaux appels de donnée, nous réservons des canaux pour les nouveaux appels voix et les appels handover.

Un problème majeur dans les réseaux mobiles est la réduction du nombre d'appels forcés à la terminaison suite à des échecs de handover. Afin d'établir la priorité des appels handover sur les nouveaux appels, notre système système applique la technique de réservation des canaux de garde et la technique de mise en attente.

La technique de réservation des canaux de garde consiste à réserver des canaux exclusivement pour les requêtes handover. D'autre part, la technique de mise en attente repose sur l'existence d'un intervalle de dégradation pendant lequel le mobile traverse la “HO area”; cette dernière est la région d'entrelacement de deux cellules voisines. Si pendant cet intervalle de dégradation, la requête handover émise par le mobile ne trouve pas de canaux disponibles dans la cellule voisine, alors cette requête est insérée dans une file d'attente.

Nous examinons la performance de deux techniques utilisées pour ordonnancer les requêtes handover voix (HO-voice) et les requêtes handover de donnée (HO-data) : “Head of the line " (HOL) et “Queue Length Threshold” (QLT). HOL consiste à servir la file de donnée si la file des appels voix est vide. Quant à QLT, elle donne la priorité aux appels de donnée lorsque la file de donnée atteint un certain seuil.

Dans un premier temps, nous supposons que la longueur du trafic de donnée est exponentielle et nous modélisons analytiquement le système. Afin de rendre le modèle analytique associé à notre système proposé tractable, nous adoptons des hypothèse Markoviennes. Ainsi, nous modélisons l’état d’une cellule par une chaîne de Markov à temps continu et à états discrets à trois dimensions. Nous résolvons alors les équations de Kolmogorov et nous sortons les paramètres de performance que nous calculons d’une manière exacte à l’aide d’un programme que nous avons codé en Maple VI interfacé avec Matlab.

L’analyse des performances prouve que le schéma proposé avec QLT améliore la qualité de service des appels de donnée sans pour autant induire une dégradation perceptible de la qualité de service de la voix. En effet, le temps d’occupation d’un canal par un appel de donnée étant nettement inférieur à celui d’un appel voix, les appels voix arrivent à être servis sans pour autant être dégradés; et ceci même si la priorité est affectée aux appels de donnée. Nos paramètres de performance sont évalués avec différentes valeurs de canaux de garde. Nous trouvons alors que les meilleurs résultats sont obtenus lorsque le nombre de canaux de garde réservés aux appels handover est plus petit et le nombre de canaux réservés aux nouveaux appels voix et aux appels handover est plus grand.

Dans le but de valider le modèle analytique, nous avons élaboré une simulation faite avec OMNeT++ [OMN]. Les résultats du modèle analytique sont très concordants avec ceux de la simulation. Ce qui prouve l’exactitude de notre modèle analytique.

Dans un second temps, nous généralisons notre travail à des trafics plus complexes comme ceux issus d’une session Web. Deux types d’allocation de canaux sont considérés: le modèle de circuit pour la voix et l’allocation par rafales pour les données. Les résultats de cette étude généralisée viennent appuyer les résultats trouvés auparavant: QLT améliore la performance des appels de donnée sans induire une dégradation perceptible à la qualité de service des appels voix.

Proposition d’une architecture dynamiquement adaptable DYNAA

Contexte

Notre seconde contribution poursuit les études faites pour la première contribution en intégrant des applications pouvant s’adapter au réseau.

La fourniture de la qualité de service présente d’importants défis techniques à relever. Un défi majeur serait la fluctuation des ressources dans le réseau. Notons que la fluctuation dans la disponibilité des ressources est principalement dûe à la mobilité. Ce problème est encore plus présent dans le réseaux ad hoc qui ne présente pas d’infrastructure et où tous les noeuds de ce réseau bougent.

L’adaptabilité est une approche intéressante qui pourrait palier au problème de fluctuations présent dans les réseaux mobiles. Cette approche établit une collaboration entre les applications d'une part et le réseau d'une autre part afin d'adapter les besoins de l'application à l'état du réseau tout en respectant le contrat de service établi avec l'utilisateur.

Proposition et Résultats

Nous proposons une architecture appelée DYNAA dont le but est d'introduire l'adaptabilité dans les réseaux cellulaires. Avec cette approche, les applications multimédias sont adaptables; ainsi elles sont capables de prendre en compte l'état du réseau. Le réseau doit offrir l'adaptabilité en remontant certaines informations à l'application et en appliquant des algorithmes d'adaptation de bande passante. En d'autres termes, grâce à l'adaptabilité la QoS de bout en bout est élaborée conjointement entre l'application et le réseau afin de satisfaire l'utilisateur de manière à respecter son contrat de service.

D'un autre côté, nous voulons atteindre les deux objectifs suivants:

1. Nous voulons empêcher la fréquente adaptation des appels tout en optimisant les performances du réseau. En effet, l'adaptation permet d'accepter un nombre plus grand de nouveaux appels et d'appels handover. Ce qui impliquerait une diminution de la probabilité de terminaison forcée de l'appel et une meilleure performance du réseau. Cependant, une fréquente adaptation des appels et une fluctuation fréquente du niveau de qualité de service sont mal perçues par l'utilisateur. Par conséquent, nous visons simultanément une adaptation moins fréquente des appels et une optimisation des performances du réseau.

2. Nous préconisions une approche qui adapte dynamiquement les appels tout en prenant en compte la charge de la cellule et la mobilité.

Dans cet esprit, nous étendons le contrôle d'admission déjà proposé en intégrant deux classes de service qui gèrent des applications présentant deux profils d'adaptabilité différents.

Nous proposons alors un contrôle d'admission et des algorithmes d'adaptation de ressources qui adaptent dynamiquement les appels tout en prenant en compte la mobilité et la charge de la cellule.

Avec la notion d'adaptabilité, deux nouveaux paramètres de QoS apparaissent et méritent d'être cités. Ce sont le "Degradation Degree" (DD) et le "Degradation Ratio" (DR). Alors que DD désigne le degré de dégradation, DR désigne la fréquence de dégradation.
Afin d'évaluer les performances du système, nous avons alors simulé notre architecture avec plusieurs scénarios. Les résultats montrent que notre approche améliore sensiblement la qualité de service en terme de probabilité de blocage des nouveaux appels, de probabilité de terminaison forcée des appels et du délai d'attente dans les files pour les requêtes handover. Cette nette amélioration de performance s'accompagne de l'augmentation des paramètres de dégradation. Ainsi, il faudrait choisir les paramètres convenables de dégradation et dimensionner le réseau en fonction des paramètres adoptés.

Etude de handover et optimisation du routage dans le sens montant dans les réseaux Cellular IPv6

Contexte

Les réseaux publics d'accès mobile, comme UMTS sont en train d'évoluer rapidement vers une définition de réseau entièrement IP. Énormément de recherche est faite dans ce domaine pour enrichir IP des fonctionnalités nécessaires pour gérer la mobilité tout en conservant la simplicité et la flexibilité qui ont fait son succès. Une amélioration célèbre est Mobile IP qui définit un protocole permettant aux stations IP de changer leur point d'attache dans le réseau.

Malheureusement, les mécanismes de Mobile IP présentent certains inconvénients, notamment un important temps de latence. Pour adresser ces problèmes, on divise généralement la gestion de la mobilité en deux parties: la micro-mobilité et la macro-mobilité.

Alors que la macro-mobilité concerne les mouvements des utilisateurs à grande échelle, la micro-mobilité désigne les mouvements des mobiles à petite échelle. Le protocole Mobile IP est bien adapté pour gérer l'itinérance et par conséquent est utilisé dans les réseaux de macro-mobilité. Pour les réseaux de micro-mobilité, plusieurs protocoles ont été définis. Les protocoles de micro-mobilité n'ont nullement l'ambition de substituer Mobile IP; ils complètent Mobile IP.

Cellular IPv6 est un protocole de micro-mobilité qui offre entre autres une gestion efficace de localisation, la connectivité passive permettant de joindre les mobiles inactifs, le support du handover proactif. Malheureusement, le protocole Cellular IPv6 souffre du routage non optimisé dans le sens montant pour un trafic entre deux mobiles d'un même réseau et du manque de support de qualité de service.

Dans une première étape, nous mettons l'accent sur les problèmes de handover dans les réseaux Cellular IPv6 et nous proposons une amélioration du protocole de routage. Dans une seconde étape, nous nous penchons sur la qualité de service dans les réseaux Cellular IP.

Proposition et Résultats

Comme nous l'avons souligné, le protocole Cellular IPv6 [SGCW00] souffre du manque d'optimisation du routage dans le sens montant pour un "intra-network traffic". Ce dernier est le trafic échangé entre deux mobiles localisés dans le même réseau Cellular IPv6. En effet avec le protocole Cellular IPv6, ce trafic devrait toujours passer par la passerelle du réseau avant d'arriver à destination. Ceci est valable même pour deux mobiles localisés dans une même cellule.

Ce type de routage pourrait alors augmenter le délai et la gigue et pourrait également impliquer une perte de bande passante. Afin de traiter ce problème, nous avons proposé et défini un mécanisme permettant l'optimisation de ce type de routage.

Par ailleurs, la perte de paquets durant le handover devrait être minimal. Afin de minimiser cette perte de paquets et améliorer la performances du handover, nous avons appliqué le mécanisme de "buffering" qui permet de garder les paquets dupliqués dans des files d'attente localisées au noeud le plus proche de la future cellule du mobile en déplacement. Après l'établissement du handover, les paquets en question seront envoyés suivant la route optimale.

L'analyse du mécanisme proposé a été menée au moyen d'une simulation. Les résultats prouvent une diminution en moyenne du délai des paquets, du nombre de noeuds traversés par le trafic, de la charge de signalisation et de la charge au niveau de la passerelle. Ces résultats montrent que notre mécanisme allège la charge au niveau de l'entité intelligente du réseau qui est la passerelle, d'optimiser le routage tout en améliorant la performance du handover.

Etude de Qualité de Service de bout-en-bout dans les réseaux Cellular IPv6

Contexte

Cette contribution vient compléter la contribution précédente en se focalisant sur la gestion de la qualité de service dans les réseaux Cellular IPv6. Nous avons remarqué lors de notre étude du protocole Cellular IPv6 que la passerelle représente le véritable goulet d'étranglement. En effet, la passerelle constitue l'entité intelligente responsable des décisions concernant l'admission, le filtrage et l'acheminenment des appels. D'un autre côté, tous les paquets devraient passer par la passerelle avant d'être acheminés vers leur destination.

Dans un souci d'alléger la charge au niveau de la passerelle et de distribuer l'intelligence dans le réseau, nous avons réalisé une étude de qualité de service dans les réseaux Cellular IPv6. L'étude est menée au moyen de l'architecture DYNAA déjà proposée.

Proposition et Résultats

Nous proposons d'ajouter des entités intelligentes, nommées DCs, distribuées dans les cellules responsables du contrôle d'admission au niveau des cellules. Nous intégrons également un champ de contrôle dans les paquets IP permettant de tenir en compte l'adaptabilité de l'architecture DYNAA.

Nous comparons entre trois types de contrôles d'admission (CAC); un CAC centralisé, un CAC distribué et un autre hybride qui est centralisé et distribué à la fois.

L'analyse des performances montre que les meilleurs résultats en termes de probabilité de terminaison forcée, de paramètres de dégradation et de probabilité de blocage des nouveaux appels appartenant à la classe prioritaire, sont obtenus avec le schémas hybride. Le schémas distribué présente de bons résultats de performance. Puisqu'il n'induit pas une charge de signalisation liée au contrôle d'admission entre la passerelle et les DCs, le schémas distribué pourrait être intéressant en cas de forte charge et pour un trafic sporadique. Il faudrait alors faire un compromis entre le schémas hybride et le schémas distribué.

Dans une étape ultérieure, nous avons complété notre étude en couplant une utilisation du protocole de type DiffServ dans le réseau coeur et IntServ dans le réseau d'accès. Cette étude constitue une approche de l'appréhension de la qualité de service de bout en bout.

P3-DCF: Différentiation de Service dans les réseaux IEEE 802.11 WLANs

Contexte

Une autre contribution dans notre thèse concerne la différentiation de service dans les réseaux locaux sans fil IEEE 802.11. De très nombreuses études se sont focalisées sur la conception de ces réseaux sans fil. Il est à noter que ces réseaux souffrent du manque du support de la qualité de service.

Le mécanisme d'accès au médium le plus étudié pour les réseaux 802.11 est le mode "Distributed Coordination Function" (DCF). DCF repose sur les méthodes de contention au médium et par conséquent n'offre aucune garantie de délai et de gigue pour le trafic temps-réel.

Dans un effort d'introduire de la différentiation de service dans les réseaux IEEE 802.11, nous étudions la couche MAC définie dans ces réseaux et proposons un mécanisme permettant d'améliorer la qualité de service dans les réseaux locaux sans fil.

Proposition et Résultats

Dans cette étude, nous proposons un nouveau mécanisme, appelé P3-DCF, qui complète le protocole DCF du standard.

P3-DCF repose sur le calcul du délai d'expiration des paquets et permet d'envoyer les paquets les plus urgents en premier lieu. Ce mécanisme est vu comme étant une implémentation de la discipline Earliest Deadline First (EDF).

En outre, le mécanisme ainsi proposé introduit la notion de priorité qui varie avec le temps [Kle76]. En effet, un paquet appartenant à une classe moins prioritaire qu'une autre classe, pourrait être servi dans un intervalle de temps après lequel, les paquets appartenant à la classe prioritaire auront plus de chance pour accéder au médium. Cette notion de priorité est réalisée grâce à une fonction mathématique que nous appelons $P3_j(t)$. Cette fonction est introduite dans le calcul de DIFS qui est un intervalle de temps contrôlant l'accès au médium.

L'analyse des performances de notre mécanisme a été faite grâce à une simulation effectuée avec Network Simulator (NS) [Net]. Nous comparons notre mécanisme à un autre mécanisme proposé dans la littérature. Les résultats montrent que notre mécanisme apporte une meilleure différentiation de service, une baisse du délai moyen de bout en bout et une diminution importante de la gigue. Il est à noter que P3-DCF est élaboré d'une manière distribuée et donc possède l'avantage d'un faible niveau de signalisation.

Etude de la qualité de service dans les réseaux UMTS

Contexte

Dans les contributions déjà traitées, nous avons soulevé les problèmes de qualité de service dans les réseaux IP sans fil. Un autre type de réseaux d'accès présent dans la future génération des réseaux mobiles sera les réseaux de troisième génération, UMTS.

Après la Release 99 des réseaux UMTS basée sur la technologie ATM [3rd], les réseaux mobiles vont utiliser la technologie IP. Dans les réseaux d'accès UMTS, la couche MAC a été spécifiée afin de fournir la qualité de service aux utilisateurs mobiles. La qualité de service offerte par la couche MAC va être basée sur la qualité de service fournie par la couche IP.

Il est important d'étudier l'impact de la couche MAC sur le trafic généré par les utilisateurs d'une cellule. En effet, cette étude permet de caractériser le trafic agrégé au niveaux des routeurs de bord. Ceci est indispensable pour l'évaluation des performances dans le réseau coeur.

A la sortie de la couche MAC, le trafic voix ne devra pas être modifié puisque sa génération est synchronisée avec la couche MAC. Tandis que le trafic de données, comme le trafic Web, va être lissé à cause de sa sporadicité.

Pour simuler le trafic sortant de la couche MAC, il est indispensable d'étudier le protocole Radio Resource Control (RRC) puisqu'il a une vision globale de tous les utilisateurs et peut se rendre compte à un moment donné du risque de saturation. Dans ce cas, RRC peut reconfigurer la couche MAC afin qu'elle réduise les débits des utilisateurs actifs.

Proposition et Résultats

Dans cette étude, nous avons simulé une cellule du réseau UMTS. Nous avons considéré deux types de trafic: voix et Web. Les modèles de trafic sont ceux proposés par ETSI [Eur98], à savoir ON/OFF pour la voix et le modèle de séquence de rafales pour le trafic Web.

Nous avons intégré un ordonnanceur qui traite en priorité le trafic voix. Ensuite, les paquets de donnée Web sont ordonnancés suivant la politique Earliest Deadline First. Cette dernière préconise d'envoyer les paquets les plus urgents en terme de délai en premier.

Nous avons analysé l'impact du modèle Web proposé par ETSI sur les performances de la couche MAC. Nous avons montré que la valeur maximale permise à la longueur de paquet ne permet pas de garantir la qualité de service aux deux types de trafic, en terme de délai. Nous avons alors proposé d'adopter une valeur de 1502 octets pour la longueur maximale des paquets. Nous avons montré que cette valeur convient aux réseaux UMTS.

Afin d'améliorer la performance des données, nous avons alors étudié deux stratégies. La première stratégie consiste à jeter les paquets dont le délai a expiré. Avec cette stratégie, les délais de donnée sont réduits avec une légère augmentation de la probabilité de perte des paquets.

La seconde stratégie consiste en une allocation de ressources plus dynamique: elle suppose un intervalle de temps de transmission égal à 20 ms et une taille d'un transport block, transmis par la couche MAC, égal à 40 octets. Nous avons trouvé que la performance des données est améliorée. Cette amélioration est obtenue puisqu' une allocation de ressources dynamique au niveau de la couche MAC permet d'avoir une efficace allocation de bande passante.

Mobility Handling and Resource Allocation in Wireless Multiservice Networks

Abstract

Wireless mobile networks have witnessed the breakthrough advances of wireless communication technology and the growing interest in multimedia applications over the past decades. Within such networks, the issue of providing an efficient resource allocation scheme has come to the fore. A suitable resource allocation in the wireless multiservice networks is expected to make efficient use of the scarce wireless resource while supporting different services with different Quality of Service (QoS) metrics.

Mobility handling is another challenging issue that must be addressed in wireless multiservice mobile networks. Mobility protocols have to be designed in such a way that they prevent the forced termination of a call and that they keep the applications running transparently to the user's movement.

To this end, we oriented our efforts towards the area of mobility handling and resource allocation in wireless multiservice networks. Six main objectives have been accomplished throughout the duration of this thesis. The first two objectives concern the wireless multiservice mobile networks in general. The third and fourth objectives deal with the wireless IP networks. The fifth objective concerns the IEEE 802.11 WLAN. Finally the sixth objective deals with a QoS study in UMTS.

This dissertation report starts with an introduction to our design principles and to the ideas behind our work. The report continues with a bibliographical survey of the mobility issues and of the IP mobility protocols. We then introduce a call admission control (CAC) scheme for wireless multiservice mobile networks. The proposed CAC is a three-priority level scheme that serves four classes of service. An analytical model is developed to evaluate the performance of the proposed scheme. Simulations were carried out in order to validate the analytical model and to generalize the adopted assumptions. Performance analysis shows that the proposed scheme with the Queue Length Threshold *QLT* scheduling discipline improves the data performance without inducing a perceptible degradation of the voice QoS.

An innovative Dynamic Adaptive Architecture DYNAA is afterwards presented. Within this architecture, we introduce the adaptation concept that implies the ability of the applica-

tion to adapt to the resource fluctuation and to share with the network the QoS responsibility. By introducing further enhancement concerning the handover requests handling, the enhanced DYNAA_Wait version was studied. The behavior and performance of our architecture are investigated by simulation. Presented results show the interest of our approach.

We pursue our studies and bring the focus on the IP Wireless networks. In a first step, we present an enhanced uplink routing mechanism coupled with a smooth handover study in a Cellular IPv6 network. Our mechanism minimizes the delay, the loss experienced by the data packets during communications and especially during handover. It also reduces the signaling load on the network gateway. Simulation results show the good performance obtained at the expense of some complexity added to the Cellular IPv6 nodes.

In a second step, the resource allocation and the call admission control aspects are treated in Cellular IPv6 networks. The DYNAA architecture with its enhanced version paves the way for a possible CAC and resource allocation that can be applied in Cellular IPv6 networks. Three types of CAC were compared, according to whether they are distributed in the cells, centralized or hybrid. We then investigate the issue of providing an end-to-end quality of service. Thus, we proposed an end-to-end QoS approach using *IntServ* in the Cellular IPv6 networks and *DiffServ* in the core network.

Another contribution of this thesis focuses on the Quality of Service in IEEE 802.11 WLAN. We propose a novel mechanism within the MAC sub-layer called P3-DCF. The proposed mechanism is elaborated in a distributed manner without any central control, thus reducing the signaling load and the transmission overheads. It further establishes a per-flow differentiation and manages to schedule the packets with an Earliest Deadline First discipline. The performance evaluation of P3-DCF is promising and demonstrates that it enhances the differentiation among flows when compared to other approaches.

The last contribution of this thesis deals with a QoS study of the MAC layer in UMTS. This study allows us to characterize the aggregated users' flows at the edge routers. This is an important task in the perspective of the performance evaluation of the core network.

We have analyzed the impact of the ETSI Web traffic model on the MAC performance. We showed that the value for the maximum allowed packet size is inappropriate in order to guarantee the required QoS (data delay). We proposed to adopt a smaller maximum size of data packets equal to 1502 bytes. We showed that this maximum size is appropriate for the UMTS.

In order to further improve the data performance, we have studied two strategies. The first strategy consists to drop the packet that has an expired delay. With this strategy, the data delays are reduced with a slight increase in the loss probability. With the second strategy that leads to a dynamic resource allocation on the MAC layer (i.e. that supposes having a time transmission interval of 20 ms and a transport block size of 40 bytes), the data performance is improved as well. This improvement is achieved because a dynamic resource allocation on the MAC layer leads to an efficient bandwidth allocation.

Contents

List of Figures

List of Tables

List of Abbreviations

AP	Access Point
BSA	Basic Service Area
BSS	Basic service Set
CAC	Call Admission Control
CBR	Constant Bit Rate
CFP	Contention Free Period
CIP	Cellular IPv6
CN	Corresponding Node
CoA	Care-of Address
CP	Contention Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CTS	Clear To Send
DCF	Distributed Coordination Function
DiffServ	Differentiated Services
DYNAA	Dynamic Adaptive Architecture
EDF	Earliest Deadline First
FA	Foreign Agent
FTP	File Transfer Protocol
HA	Home Agent
HO	Handover, Handoff
HOL	Head Of the Line
IntServ	Integrated Services
MAC	Medium Access Control
MN	Mobile Node
NAV	Network Allocation Vector
P3-DCF	Per-Packet Priority Using DCF
PCF	Point Coordination Function
QLT	Queue Length Threshold
QoS	Quality of Service
RSVP	Reservation Protocol
RTS	Request To Send
SLA	Service Level Agreement
TCP	Transmission Control Protocol
UDD	Unconstrained Delay Data
UDP	User Datagram Protocol
WLAN	Wireless Local Area Network

Chapter 1

Introduction

The mobile cellular era started with the *first generation mobile systems*. These systems used analog transmission for speech services. Mobile communications have undergone significant changes and experienced enormous growth since then ...

Second generation systems using digital transmission were then introduced. They offer data services, higher spectrum utilization efficiency and more advanced roaming than the first generation systems. The services offered by these systems cover speech and low bit rate data. The second generation systems further evolved towards systems commonly referred to as *generation 2.5*. This generation offers more advanced services for circuit and packet switched data than the second generation systems.

Third generation systems with the Universal Mobile Telecommunication Systems (UMTS) will offer high rate data services, high spectrum utilization efficiency, simultaneous multiple services for one user, and services with different quality of service classes.

At the moment, a remarkable interest for wireless local area network (WLAN) is observed. WLANs are appropriate to cover temporary events such as forums and conferences. Lastly, operators plan to use them as systems able to provide Internet access in public or private locations that are called "hot-spots" which include airport, railway stations and hotels for example.

Even it is difficult to predict the long term situation, especially in the mobile telecommunication field, it can be assessed that the next generation of mobile networks, currently designed by *Beyond 3G* or *fourth generation systems* will be built to transparently support both mobile/fixed and wireless/wired access environments.

Beyond 3G is an evolutionary system that progressively incorporates heterogeneous wireless access technologies and supports seamless IP-based mobile multimedia services. This leads to the implementation of multi-interface user terminals, capable of handling vertical handovers between access networks based on different technologies. Furthermore, this leads to the evolution of wireless system architectures, such as cellular, broadcast, WLAN and satellite technologies. On the other hand, the IP networking technologies must be enhanced in order to enable a seamless mobile Internet, incorporating existing/emerging and wired/wireless access technologies.

With the constant improvement of wireless technology and the explosive growth of wireless communication market, the demand for newer multimedia applications is increasing rapidly. The convergence of wireless technology and multimedia application presents network operators with enormous opportunities as well as great challenges. *Quality of Service (QoS) Provisioning* and *Mobility Management* are two key challenging issues that must be addressed in wireless multiservice mobile networks.

QoS provisioning in wireless mobile networks supporting multimedia applications have to meet the expectations of users while maintaining reasonably high utilization of radio resources. The QoS provisioning problem is more challenging than in fixed networks for two main reasons.

First, the link bandwidth resource is limited in a wireless environment. Second the changing environment in wireless networks due to the user's mobility and interference results in varying bandwidth. Thus, how to allocate and how to use the limited wireless resources efficiently are to be studied.

On the other hand, the mobility management is another important issue to be addressed when studying the wireless mobile networks. A mobile user must be able to freely move across the networks while maintaining its current communications. An interrupted communication is a very frustrating phenomenon that may happen to a user. Thus, an efficient mobility protocol must manage to avoid the forced termination of an ongoing call.

The mobility management is not an easy task to achieve. This task is even more complicated in wireless IP networks. In these networks, the IP address of a user must change when the mobile user hands off to a network. Thus, mobility protocols have to be designed in such a way that they prevent the forced termination of a call and that they keep the applications running transparently to the user's movement.

The goal of our thesis is to study the mobility issues and the resource allocation in a QoS framework within multiservice wireless mobile access networks. The studied networks regroup second, third and fourth generation of mobile networks.

1.1 Context of the thesis

This thesis work is performed in the context of two projects: the RNRT Minicell project and the ITEA Ambience project.

1.1.1 Thesis Objectives

Our thesis focuses on the QoS provisioning in wireless multiservice networks. Before identifying our objectives, we give an abstract view of a QoS framework [G.P]. Figure 1.1 shows three layers that try to provide QoS to the users. These layers are the "business layer", the "network layer" and the "physical layer".

At the business layer, competing service providers offer services to clients who may freely choose among the service types and service providers. Before the service provision, a client needs to negotiate *a service contract* that determines the contractual constraints for the service.

Part of the service contract will be an agreement on the quality of service that is expected by the client and that must be guaranteed by the service provider. Obviously, QoS parameters are application specific and their actual values depend on the current state of the whole distributed system, e.g. on the load condition of the servers, on the network throughput, and so on.

In order to provide the expected QoS to the client, the service provider must ensure that this QoS is mapped to the "network layer" and to the "physical layer" (figure 1.1). The network layer provides the QoS to the different services through efficient methods of resource management, call admission control, bandwidth adaptation algorithms, ...

The QoS provided at the network layer must be mapped to the physical layer that provides efficient scheduling disciplines, policing, buffering, . . .

As we can see, the QoS is not specific to a layer. It must be present in the different layers in order to provide an end-to-end QoS to the client as defined in the service contract.

Our thesis focuses on the QoS provisioning at the network layer through the design of efficient call admission control, of radio resource management, and of the medium access control (MAC), while taking into consideration the users' mobility.

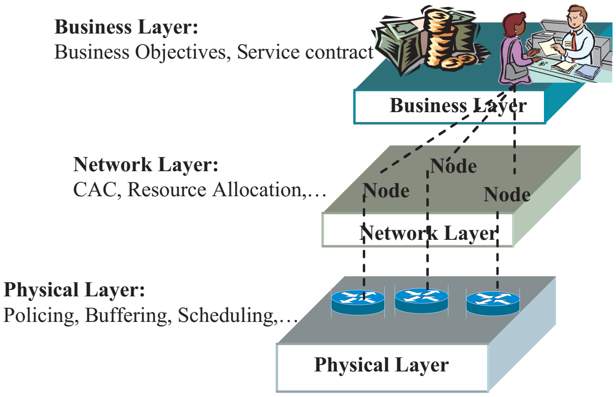


Figure 1.1: QoS Mapping on the Different Layers

Our work consists of the following design principles:

- The maximum utilization of the scarce uplink resources is to be realized. The narrow bandwidth of a wireless network is a bottleneck and should be well dimensioned.
- The fluctuation in resource availability in wireless and mobile network is much more severe than the level of fluctuation in wired networks and this is mainly due to the users' mobility. In our thesis, we dimension the wireless network. We do not deal with the core network issues and we focus on the access network.
- The user's call should encounter minimum disruption. In wireless network admission, a decision is made on new and handover calls. Since forced call terminations due to handover failure have significant negative impacts on the user's perception of network reliability, handover requires special considerations. We strive towards a solution that provides the minimum forced termination probability due to the user's mobility.
- We believe that there must be a fundamental change in the expectations we have from QoS provisioning. The end-to-end QoS provisioning is no longer the sole responsibility of the network nor of the application. Now, this responsibility is shared between the application and the network.

  Multimedia applications need to be adaptive, to renegotiate the service request and to deal with changing conditions. End systems must be network aware as they must take into account the network status and must be able to adapt the multimedia streams accordingly. On the other hand, the network must provide different levels of QoS to the mobile users.

This adaptive approach implies that the end-to-end QoS provisioning is shared between the application and the network. They must together choose the QoS level to deliver multimedia content to a wireless mobile terminal in the most acceptable form given the resources in the network.

The adaptive approach is revealed by the new trend towards "intelligent" terminals. The new terminals are becoming more sophisticated and equipped with advanced technologies enabling them to interact with the network.

1.2 Thesis Contributions

Six main objectives have been accomplished throughout the duration of this thesis. The first two objectives concern the wireless multiservice mobile networks in general. The third and fourth objectives deal with the wireless IP networks. The fifth objective concerns the IEEE 802.11 WLAN. Finally the sixth objective deals with a QoS study in UMTS.

To be more precise,

1. The first objective concerns the design of a multiservice call admission control supporting four classes of service in the wireless multiservice mobile networks.

2. The second objective aims at enhancing the proposed call admission control. This objective focuses on developing a dynamic adaptive architecture DYNAA that establishes a collaboration between the application and the network. This collaboration handles the user's mobility and the high variability in network conditions while offering the best possible service to the user.

3. Because next generation wireless networks will be most probably IP-based and are expected to inter-work with the Internet backbone seamlessly, we propose to study the "Cellular IPv6" protocol. Two major drawbacks of this protocol are support of an optimized uplink routing mechanism and provision of quality of service. These will be the subject of the third and fourth achievement.

   The third achievement is then the study of a smooth and anticipated handover along with the enhancement of the uplink routing in Cellular IPv6 networks.

4. The fourth achievement concerns the support of QoS in the Cellular IPv6. Thus, we integrate our proposed architecture DYNAA into the Cellular IPv6 protocol.

5. The fifth objective concerns the design of a novel mechanism within the MAC sub-layer, called P3-DCF. This mechanism employs distributed coordination function (DCF) as a fundamental access for prioritized services in IEEE 802.11 WLAN.

6. Finally, we made a QoS study of the MAC layer in UMTS. We have studied the Web traffic model and managed to choose the best parameters in order to improve the system performance. Further performance improvements were achieved by two proposed mechanisms. Simulations show the results obtained with our mechanisms.

1.3 Thesis Organization

The following chapter of this report starts with a bibliographical survey of the mobility and handover issues proposed for wireless multiservice mobile networks. After a discussion of the mobility issues and handover challenges, we expose the main handover priority techniques. Then, we present the hierarchical cellular systems that improve the handover handling.

The mobility protocols in IP-based networks are exposed in Chapter 3. In a first step, we present the problems that an IP mobility protocol must face. Then, we describe the "Mobile IP" which supports the global Internet mobility. Next, we expose the "IP Micro-mobility protocols" that are designed in order to limit the disruption to user traffic during handover and to handle frequent handovers across multiple subnetworks. Afterwards, we make a comparative study between the different micro-mobility protocols, considering some mobility management issues.

Chapter 4 and Chapter 5 are dedicated to the study of the QoS and mobility issues in wireless multiservice mobile networks. They concern the work done for the first and second objectives.

In Chapter 4, we propose an efficient call admission control (CAC) scheme for wireless multiservice mobile networks. The proposed CAC is a three-priority level scheme that serves four classes of service. Two types of scheduling are implemented in order to serve the handover requests. In a first step, we use the analytical approach in order to evaluate the performance of the proposed scheme: a three-dimensional Markov chain is elaborated in order to model the system. In a second step, we validate the analytical model by a simulation. Finally, we generalize our assumptions by taking a Web session model for data traffic. Two types of channel allocation are considered: dedicated for voice and burst allocation for data. The performance analysis of the two scheduling schemes is discussed.

The ability of the application to adapt to the resource fluctuation and to share with the network the QoS responsibility is the main subject of the work presented in Chapter 5. This chapter defines a dynamic adaptive architecture DYNAA. The proposed architecture focuses on the collaboration between the application and the network in order to handle the user's mobility and the high variability in network conditions while offering the best possible service to the user.

DYNAA defines an admission control based on adaptive bandwidth algorithms in order to provide QoS to two classes of service. It is shown that, by dynamically adjusting the allocated bandwidth while taking into account the current network conditions, our proposed architecture can be dynamic and consequently achieve better QoS. The results obtained by simulations illustrate the enhanced characteristics of the proposed architecture.

Wireless access to the Internet may outstrip all other forms of access in the near future. Therefore, a significant part of this thesis work is oriented toward the study of the QoS and the mobility handling in Wireless IP networks. Chapter 6 and Chapter 7 present this part of the work. They concern the work done for the third and fourth objectives.

Cellular IPv6 is a promising micro-mobility protocol that presents some important features, such as a cheap passive connectivity, an efficient location management, an efficient routing and a flexible handover. Nevertheless, the Cellular IPv6 protocol lacks two important issues: an optimized uplink routing and a QoS support. In Chapter 6, we address the uplink routing issues. In Chapter 7, a QoS study is carried out.

In Chapter 6, we study the uplink routing for intra-network traffic in Cellular IPv6 networks.

We propose an enhancement for this mechanism. Secondly, a smooth and anticipated handover is presented. Simulations and performance analysis are presented before concluding the chapter.

In an attempt to provide QoS for users generating inter-network traffic, we investigate in Chapter 7 the issue of providing an efficient CAC that offers QoS to different classes of service. Therefore, we apply the proposed architecture DYNAA in the Cellular IPv6 networks. The simulations carried out intend to evaluate the performance of three types of CAC: a centralized CAC located in the network gateway, a CAC distributed in the cells without any central coordination, and a hybrid CAC. This chapter continues with the proposal of an end-to-end QoS framework with Cellular IPv6 in the micro-mobility domains.

Research efforts are oriented towards the enhancement of the MAC in IEEE 802.11 WLAN. Providing differentiated services in a WLAN requires that the MAC supports some degree of separation between different types of services. In the literature, several approaches compete for providing quality of service enhancements in IEEE 802.11. Among others, some approaches investigate differentiated services by extending the Distributed Coordination Function (DCF). These approaches are exposed in Chapter 8.

Based on the DCF, we propose a novel mechanism within the MAC called P3-DCF. An enhanced function, Per Packet Priority (P3), integrated to the DCF establishes not only a per-flow differentiation but manages to schedule the packets with an Earliest Deadline First discipline as well. Furthermore, the proposed mechanism is elaborated in a distributed manner without any central control, thus reducing the signaling load and the transmission overheads.

The simulations carried out show that the proposed mechanism satisfies the maximum tolerable latency and jitter bounds of real-time traffics and performs efficient flow differentiation. The performance evaluation of P3-DCF is promising and demonstrates that it enhances the differentiation among flows and the delays experienced by each flow when compared to other approaches.

Chapter 9 deals with a QoS study of the MAC layer in UMTS. This study allows us to characterize the aggregated users' flows at the edge routers. This is an important task in the perspective of the performance evaluation of the core network.

We have analyzed the impact of the ETSI Web traffic model on the MAC performance. We showed that the value for the maximum allowed packet size is inappropriate in order to guarantee the required QoS (data delay). We proposed to adopt a smaller maximum size of data packets equal to 1502 bytes. We showed that this maximum size is appropriate for the UMTS.

In order to further improve the data performance, we have studied two strategies. The first strategy consists to drop the packet that has an expired delay. With this strategy, the data delays are reduced with a slight increase in the loss probability. With the second strategy that leads to a dynamic resource allocation on the MAC layer (i.e. that supposes having a time transmission interval of 20 ms and a transport block size of 40 bytes), the data performance is improved as well. This improvement is achieved because a dynamic resource allocation on the MAC layer leads to an efficient bandwidth allocation.

Finally, Chapter 10 presents a general summary of the work achieved and conclusions concerning the results obtained during this thesis. Some perspectives and open questions are given for the continuation of our work.

1.4 Before Starting

Before going into details of the work, it should be noted that the traffic modeling is important in the networks implementation. The traffic modeling should take into account the model of the traffic sources, the traffic aggregation and the traffic scenarios.

In our thesis, we have devoted a part of our work to define the models of the traffic. In this section, we present three traffic models that were chosen for the implementation and we give as well the necessary parameters related to the models [Eur98].

1.4.1 Voice Traffic Model

The voice calls are generated according to a Poisson process assuming a mean call duration of 120 seconds. The voice traffic model is an on-off model, with activity and silent periods being generated by an exponential distribution (figure 1.2). In the activity period (or talk-spurt), packets of constant size are generated at constant time intervals. In the silence period, a silence descriptor is sent at the beginning of the silence duration. Then, a silence descriptor is sent each time there is a change in the background noise. Parameters of the voice traffic are depicted in table 1.1. Note that this voice traffic model has been proposed by the 3GPP for the UMTS.

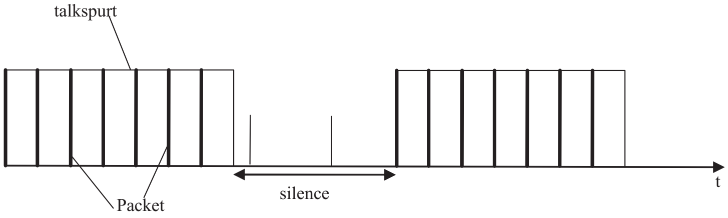


Figure 1.2: Voice Traffic

Parameter	Distribution	Mean value
Call Duration	Exponential	120s
Talk-spurt Duration	Exponential	350 ms
Silence Duration	Exponential	650 ms
Interarrival time between packets	Deterministic	20 ms
Size of silence packets	Deterministic	35 bits
Size of voice packets	Deterministic	244 bits

Table 1.1: Parameters of Voice Traffic

1.4.2 HTTP Traffic Model

Figure 1.3 depicts a typical WWW browsing session, which consists of a sequence of packet calls. The user initiates a packet call when requesting an information entity. During a packet call several packets may be generated, which means that the packet call constitutes of a bursty

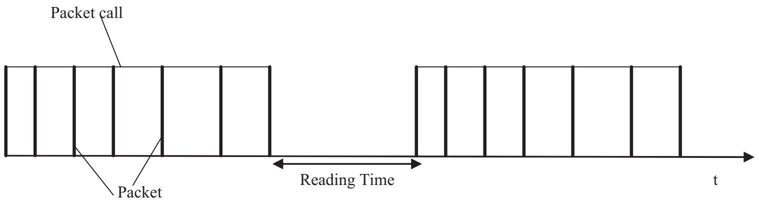


Figure 1.3: HTTP traffic

sequence of packets. The burstyness during the packet call is a characteristic feature of packet transmission in the fixed network.

A packet service session contains one or several packet calls depending on the application. After the document is entirely arrived to the terminal, the user is consuming certain amount of time for studying the information. This time interval is called reading time. The following parameters must be modeled in order to catch the typical behavior described in Figure 1.3:

- The *session arrival* process is modeled as a Poisson process.
- The *number of packet calls* per session, N_{pc}, is a geometrically distributed random variable with a mean $\mu_{N_{pc}}$ [packet calls].
- The *reading time* between packet calls, D_{pc}, is a geometrically distributed random variable with a mean $\mu_{D_{pc}}$ [model time steps].
- The *number of packets* within a packet call, N_d, is geometrically distributed random variable with a mean μ_{N_d} [packet].
- The *time interval* between two consecutive packets inside a packet call D_d is a geometrically distributed random variable with a mean μ_{D_d} [model time steps].
- The *size of a datagram*, S_d, is a random variable with a Pareto distribution with cut-off.

The packet size $PacketSize$ is defined with the following formula: $PacketSize = min(P, m)$
,
where P is normal Pareto distributed random variable ($\alpha = 1.1, k = 81.5bytes$) and m is maximum allowed packet size, $m = 64kbytes$. The cumulative distribution function of the packet size is:

$$F(x) = \begin{cases} 1 - (\frac{k}{x})^{\alpha} & k \leq x < m \\ 1 & x = m \\ 1 & x > m \end{cases}$$

The probability measure of the packet size is:
$P(dx) = \frac{k^{\alpha}}{\alpha x^{\alpha+1}} 1_{[k,m]}(x)dx + \frac{k^{\alpha}}{m^{\alpha}} \delta_m(dx)$
where

1. dx is the Lebesgue standard measure

2. δ_m is the Dirac measure

3.

$$1_{[k,m]}(x) = \begin{cases} 1 & if x \in [k,m] \\ 0 & otherwise \end{cases}$$

The average packet size μ can be easily calculated as:

$$\mu = \frac{\alpha k - m(\frac{k}{m})^{\alpha}}{\alpha - 1} \tag{1.1}$$

In the sudy done in Chapter 9, we simulated the MAC layer in UMTS. In our simulation study, we adopted the Web traffic as proposed in ETSI. We showed that the maximum allowed packet size (64 kbytes) induces bad performance in terms of delay. Thus, we proposed to reduce the maximum HTTP packet size to 1502 bytes which is the case in the local networks. The proposed maximum packet size will be adopted in the simulations carried out in the next chapters.

With the parameters above the average size is: $\mu = 287 bytes$.

Table 1.2 gives default mean values for the distributions of typical WWW service having different values for the unconstrained delay data (UDD). According to the values for α and k in the Pareto distribution, the average packet size μ is 287 bytes. Average requested file size is $\mu_{N_d}.\mu \approx 7187 bytes$. The inter-arrival time is adjusted in order to get different average bit rates at the source level.

WWW UDD	$\mu_{\mathbf{N_{pc}}}$	$\mu_{\mathbf{D_{pc}}}$	$\mu_{\mathbf{N_d}}$	$\mu_{\mathbf{D_d}}$	**Parameters for packet size distribution**
32 kbits/s	5	412	25	0.074	k= 81.5, α=1.1
64 kbits/s	5	412	25	0.037	k= 81.5, α=1.1
144 kbits/s	5	412	25	0.016	k= 81.5, α=1.1

Table 1.2: Parameters of HTTP Traffic

1.4.3 FTP Traffic Model

The FTP traffic is modeled as the HTTP traffic. However, we consider that there is one packet call per session. The parameters of the FTP traffic are shown in table 1.3.

$\mu_{\mathbf{D_{pc}}}$	$\mu_{\mathbf{N_d}}$	$\mu_{\mathbf{D_d}}$	**Parameters for packet size distribution**
30 min	200	0.037	k= 81.5, α=1.1

Table 1.3: Parameters of FTP Traffic

Chapter 2

Mobility Issues

2.1 Introduction

One major challenge in current and future wireless mobile networks is the provision of mobility which constitutes a key element in wireless communications. Therefore, special care has to be taken when handling mobility. Different types of mobility can be considered:

- User Mobility. This kind of mobility provides the possibility for a mobile user to communicate in different locations and while on-the-move. This is made possible by the fact that terminals are not attached to the fixed infrastructure by wires. This type of mobility can be characterized by three distinct features namely; the nomadism, the horizontal handover and the vertical handover.

    1. Nomadism: it allows the user to obtain access to networks from multiple access points. It includes delivery of a service to a subscriber, while roaming between different networks. When moving across networks, the mobile user expects to easily find his required services even if he changes his terminal. The network must offer a flexible and scalable framework for providing mobile users with access to information about the existence, location, and configuration of networked services.
    2. Horizontal handover: a horizontal handover is typically an intra-technology handover. It refers to the handover in which the mobile node moves across networks with the same technology.

       For IP-based networks, the horizontal handover refers to the handover in which the mobile node's network interface does not change (from the IP point of view). The mobile user communicates with the access network via the same network interface before and after the handover. For IP-based networks, a horizontal handover is typically an intra-technology handover. However, it can be an inter-technology handover as well if the mobile can do a layer 2 handover between two different technologies without changing the network interface seen by the IP layer (see Chapter 3). This kind of mobility does not need the support of the core network.
    3. Vertical handover: a vertical handover is typically an inter-technology handover. It refers to the handover in which the mobile node moves across networks with different technology.

For IP-based networks, the mobile node's network interface to the access network changes in a vertical handover. In these networks, a vertical handover is an inter-technology handover but it can be an intra-technology handover if the mobile changes its network interface (from the IP point of view). This kind of mobility needs the support of the core network.

- Mobility of the network devices: this concerns the mobility of some devices in the network. This the case of the ad hoc infrastructure-less mobile-appliance network which is a dynamically reconfigurable wireless network with no fixed infrastructure due to the nomadism and to the mobility of the mobile appliances. This is also the case of Low Earth Orbiting (LEO) and Medium Earth Orbiting (MEO) satellite systems.

In our work, we deal with the user's mobility. Thus, we refer to the user's mobility simply as mobility.

Two types of problems arise while handling the mobility issue. The first problem is the tracking of inactive terminals in order to be able to respond quickly to requests from the fixed network to establish communications with these terminals.

The second problem is that a moving active terminal will face the risk that it will leave the area where its current radio access network point is capable of providing a certain level of quality of service. This problem is of real-time character and is particularly demanding when it comes to real-time traffic, where seamless service is required with little or no loss of transmitted data is permissible. This problem concerns the handover handling.

This chapter deals with the handover issues in wireless mobile networks. The handover handling constitutes one of the most important aspects that largely determine the performance of wireless mobile networks. Since, a forced termination of an ongoing call due to the handover failure is more annoying to a user than the refusal of a new call, the handover prioritization techniques come to the fore. Section 6.6 lists the handover procedure along with the handover prioritization schemes.

On the other hand, in order to cope with terminals with different moving speeds and in order to achieve high user capacity, the "multilayer" architecture has been proposed for efficient management of radio resource. Modern resource allocation and handover procedures have to be optimized for such an architecture. Handover issues with speed-sensitive algorithms that allow the transfer of mobile stations between the different layers of the multilayer network are discussed in section 2.3.

2.2 Handover

A handover is defined as a change of the radio channel used by a mobile terminal. The new channel may be within the same cell (intra-cell handover) or in a different cell (inter-cell handover).

The handover process is initiated by the issuing of a handover request when the power received by the mobile from a neighboring cell's base station exceeds the power received from the base station of the current cell by a certain amount. This is a fixed value called the *handover threshold*.

For successful handover, a channel must be granted to the handover request before the power received by the mobile reaches the *receiver threshold*, i.e., the threshold in the received power, below which acceptable communication with the base station of the current cell is no longer possible (figure 2.1).

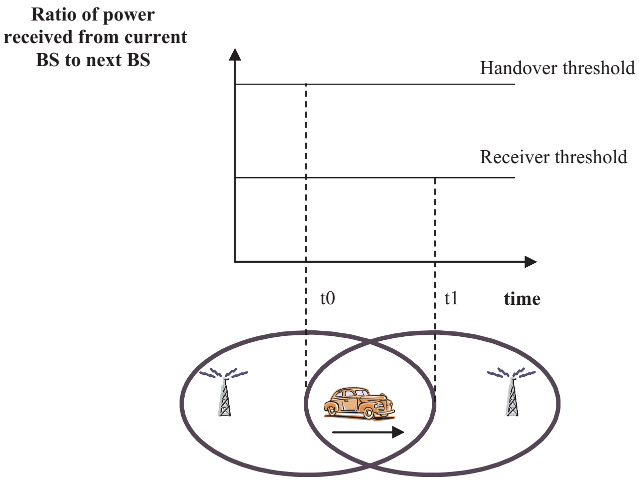


Figure 2.1: Handover Process

The *handover area* is the area where the ratio of received power levels from the current and the target base station is between the handover and the receiver thresholds. If the power level from the current base station falls below the receiver threshold prior to the mobile being assigned a channel by the target base station, the call is terminated; therefore the handover attempt fails.

2.2.1 Reasons for Handover Initiation

The main reason for initiating handover is the link quality degradation. However, handover may be initiated for other reasons. Some of the reasons are:

- Satisfying a service provider request. For example, a service provider's preferred network over which to deliver a service may be time dependent.
- Avoiding congestion. In case congestion occurs in the current cell, the terminal can be redirected to an alternative cell.
- Satisfying the required QoS. If the current cell is not capable of satisfying the QoS required by the user, the terminal can be redirected to another cell.
- Subscription (roaming). A service provider may have selected to forbid its subscribers to roam to particular networks.
- Hierarchical handover. The operators policy may be to hand off slow moving users to small cells so that the fast moving users are accepted in larger cells. This reduces the frequency of handovers suffered by the fast moving users. This is the subject of section 2.3.

2.2.2 Handover Procedure

The handover procedure can be divided into three phases [ZK01]:

1. Handover initiation/decision: in this first phase, a decision is made to initiate a handover. The decision of handover initiation is based on measurements of the link quality made by the terminal or the network. Four handover strategies have been proposed in the literature, classified depending on who initiates the handover and decides about the next base station. The four strategies are:

   - Network Controlled HandOver (NCHO): the network periodically measures the signal power in the uplink and when the signal level falls below a certain handover threshold, the network initiates a handover process. The main advantages of this method are reduced signaling load as well as low complexity of the terminal. One of the disadvantages of this method is the low reliability of the handover decision, since it is only based on radio link conditions on the uplink whereas radio link characteristics on the uplink and downlink may not be correlated.
   - Mobile Controlled HandOver (MCHO): the mobile measures the signal level on the downlink and all the signals coming from adjacent cells. Based on this information, the mobile decides to perform the handover to the base station which is detected as the best candidate for handover. This method facilitates a very fast handover and also gives more flexibility in the design of terminals. Nevertheless, it may not prove a reliable handover decision, since it is only based on the downlink radio conditions.
   - Mobile Assisted HandOver (MAHO): in this case, both the network and the mobile make measurements of radio link parameters (uplink and downlink). The downlink measurements made by the mobile are reported to the network periodically, and handover decision is made by the network based on the results of uplink and downlink measurements. The rate with which the measurements are carried out is an important parameter that has to be selected properly. If the measurements are too frequent, excessive signaling load is generated. Meanwhile, the measurements should be frequent enough to permit a rapid response when a handover is needed.
   - Network Assisted HanOver (NAHO): the decision to initiate handover is made by the mobile terminal. This decision is based on both uplink and downlink signal measurements. The network informs the mobile terminal about the uplink signal measurements. Hence, the active mobile makes handover decision with the network assistance. In this method, increased handover reliability is achieved at the cost of increased complexity of the terminal and also increased signaling load on the current radio link.

2. Handover resource assignment: when a (quality-based) decision to hand off a user has been made and a target base station has been determined, the question arises whether there are enough radio resources in the receiving base station. Thus at this stage, the network has to provide resources for the handover request.

3. Handover execution: in this last step, a target base station capable of receiving the terminal has been selected. Now, a signaling procedure to facilitate the handover is needed to inform the involved terminal and base station about the new resource allocation. This may require synchronization procedures. The signaling procedure has to be swift and reliable in order not to lose payload data while the actual switch-over is done and not to drop the connection.

2.2.3 Handover Performance

According to the handover handling techniques adopted in the network, the performance experienced by the handover may differ. A handover can be a:

- Smooth handover. In this case, the handover aims primarily to minimize packet loss, with no explicit concern for additional delays in packet forwarding.
- Fast handover. This kind of handover aims primarily to minimize delay, with no explicit interest in packet loss.
- Seamless handover. With this handover, there is no change in service capability, security, or quality. In practice, some degradation in service is to be expected.

2.2.4 Handover Prioritization Schemes

In the standard scheme known as the Non-Prioritized Scheme *NPS*, the handovers are handled in exactly the same manner as the new call arrivals [HR86]; therefore, the blocking of handover calls is the same as that of new calls.

Of particular interest in mobile systems is the reduction of the number of calls forced to terminate, because from the user and the service provider point of view, it is less desirable to force an ongoing call to terminate than to block a new call. Therefore, methods for decreasing the forced termination probability by prioritizing handovers, at the expense of a tolerable increase in call blocking probability, have been devised in order to enhance the quality of the cellular service. These are the Reserved Channel Scheme *RCS* ([HR86], [OGA99]), the Queuing Priority Schemes ([LMN94b], [TJ92], [LMN94a]) and the Subrating Scheme *SRS* ([LNH96], [LNH94]).

Reserved Channel Scheme (RCS) The simplest way to give priority to handover calls is by specially reserving channels for them in every cell. In this scheme, the available channels in a cell are divided into two sets. The first set of channels is used by new calls as well as handover calls. The second set of channels is used only by handover calls. These reserved channels are called *guard channels*. The RCS scheme provides improved performance for handover calls at the expense of a reduction in the total admitted traffic and an increase in the blocking of new calls. Another shortcoming of the employment of guard channels scheme is the risk of inefficient spectrum utilization.

Queuing Priority Schemes The queuing of handover requests is another generic prioritization scheme offering reduced probability of forced termination. Queuing of handover requests is possible because the mobile spends some period of time in the handover area, during which its communication with the current base station degrades at a rate depending on its velocity. This period of time is called the *degradation interval*. The communication degradation is easily monitored by means of radio channel measurements, usually taken by the mobile and submitted to the network. The “next handover request” to be served is selected based on the queuing discipline. In the *First-In-First Out* (FIFO) discipline, the next handover request to be serviced is the earliest one to arrive in the queue.

In the *Measurement-Based Priority Scheme* (MBPS) [TJ92], the next handover request selected is based on a non-preemptive dynamic priority policy. With MBPS, the handover area

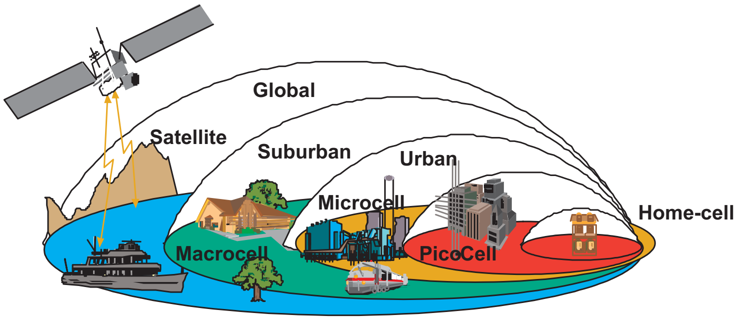


Figure 2.2: Hierarchical Networks in UMTS

can be viewed as regions marked by different ranges of values of the power ratio. These values correspond to the priority levels such that the highest priority belongs to the mobile whose power level is closest to the receiver threshold. The power levels are monitored continuously and the priority of a mobile dynamically changes depending on the power level it receives while waiting in the queue. The mobiles waiting for a channel in the handover queue are sorted continuously according to their priorities. When a channel is released, it is granted to the mobile with the highest priority. MBPS manages to decrease the forced termination probability at the cost of increasing the call blocking and of decreasing the ratio of carried traffic to offered traffic.

Subrating Scheme (SRS) In the SRS, if a base station does not have a free channel to serve a handover call, a new channel is created to serve it by subrating an existing call. Subrating means an occupied full rate channel is temporally divided into two channels at half the original rate: one serves the original call and the other serves the handover request. The probability of forced termination of existing calls and of new call attempts in this new scheme compare favorably with the standard scheme (non-prioritizing scheme) and the schemes proposed previously. However, this scheme presents an additional complexity of implementing subrating and the impact of continuing the conversation on a lower rate channel.

2.3 Handover in Hierarchical Networks

One of the major challenges in personal communications arises from the objective of serving both low and high mobility users within a system. To meet this objective, a wireless system should strive to maximize the number of subscribers while keeping the network control associated with handover at an acceptable level. To achieve the conflicting goals of maximizing network capacity (which implies use of small cells) and minimizing network control (which favors large cells), the concept of *hierarchical* or *multitier* (also called *multilayer*) networks was introduced.

With the two-tier networks concept, small cells called “microcells” are overlaid by larger cells called “macrocells”. Low mobility users (e.g., pedestrians) undergo handover only when crossing microcells boundaries, while high mobility users (e.g., people in moving cars) undergo handover only when crossing macrocell boundaries. This approach keeps the rate of handover at acceptable levels for both kinds of users. Thus macrocells must accommodate high-speed terminals whereas microcells are more adapted to low-speed terminals. Macrocells can act as overflow recipients for microcell layers. In *reversible systems*, the call may be taken back by the microcell layer as soon as a resource is available.

Third generation systems include several layers. Figure 2.2 illustrates these layers which are the home-cell, the picocell, the microcell and the macrocell. Furthermore, an additional layer may provide world-wide coverage by satellites, serving subscribers in the zones without adequate land-based coverage. It is common to associate people working in indoor environments with home-cells and picocells, pedestrians and users in cars traveling at low speeds with microcells, users in cars moving at high speeds with macrocells and users traveling in ships and airplanes with the highest level of the hierarchy or satellite.

There is no clear classification of when a user should be considered as a low-speed user or a high-speed user. The threshold to differentiate one from the other can change during the day. Thresholds are proposed by system operators to differentiate users and depend mainly on the probability density function of their speed, the time required to establish a call or a handover and the cells' radii. Successful classification of users in hierarchical networks in terms of their speed helps to reduce the forced termination of calls in progress in layers with small cells. It also reduces the signaling load between base station controllers and improves the use of channels in layers with large cells when acting as overflow servers.

Intensive deployment of microcell base stations is generally more expensive than conventional macrocell deployment because of the potential high number of sites. A hierarchical architecture has to be carefully optimized so that the layers can act in a complementary way.

Three important topics have to be studied for the hierarchical networks design:

- Determination of the best microcell area size and the position of the microcell base station inside the macrocell.
- Resource management between layers: how to share precious resources (time and frequency) between the layers is a key question. There are three main ways to allocate spectrum for different cell layers.

  1. The radio bandwidth is split between layers and each layer uses its own radio frequencies. In this case, the splitting may lead to a loss of bandwidth efficiency.
  2. The different cell types use the same instantaneous radio frequencies continuously. The transmission powers are tuned to provide acceptable carrier-to-interference C/I level.
  3. The different layers share several radio frequencies but not at the same time. Such a dynamic sharing may be provided by means of dynamic channel allocation.

  In [IGG93], the authors evaluate four approaches to sharing the spectrum between two tiers. The first two approaches feature *the spread-spectrum sharing*, i.e, they use TDMA among microcell users and CDMA among macrocell users for the first approach or vice versa for the second approach. The other two approaches feature *the orthogonal sharing*, i.e., they use TDMA in both tiers, with different time slots for the third approach and different frequency channels for the fourth approach so there is no overlap between tiers.

  The first and second approaches encounter poor capacity because of the large amounts of cross-tier interference. The capacity results for the third approach are not good since it is necessary to divide the TDMA time slots per microcell by the additional reuse factor from the macrocells.

The best approach in terms of maximizing network capacity turns out to be the one that allocates a different part of the spectrum to each tier. It has the merit of facilitating the incorporation of new systems with existing ones.

- Call and handover admission. Once the frequencies are allocated in micro and macro layers, terminals may be served by two cells belonging to different layers. The question that arises is to know which policy provides the best efficiency.

  In [Fit96], a two-layer system is considered without mobility. Three user types are distinguished: the first type of users can access only the macrocell layer and the second type of users only the microcell layer; the third type access the microcell layer first and then overflows to the macrocell if there is no available channel. The system performance is studied as a function of the proportion of the different user types. The network capacity is expressed as the offered load that provides a mean blocking probability of two percent. This capacity reaches a maximum for a given proportion of users that can access the microcell layer (type 2 and 3). If the load is high, the second type of users suffer a dramatic increase of the call blocking rate as they cannot overflow. In all cases, the maximum offered load is obtained when all calls can overflow.

  Several teletraffic analysis of non reversible hierarchical systems with overflow processes have been made [EDWS89], [SN92], and [RH94]. In [RH95], a two-layer terrestrial system is considered with an additional satellite that could provide coverage in very sparsely dense areas. Terminals are classified into high mobility and low mobility terminals. Two types of users are considered: satellite-only and dual users that can access both terrestrial and space networks. Calls are directed towards the lowest layer that gives acceptable coverage. If there is no available channel, the request overflows to the upper layer. Guard channels are considered at each layer to reduce the handover failure probability compared to the new call blocking probability. The system is found to have good performance at moderate load but degrades sharply when the offered load becomes too large.

  In reversible systems, it is advantageous to allow calls to be taken back by the microcell layer when some resources are available. As soon as a channel is released in a microcell, the system tries to take back a low mobility user. References [BMM96] and [ZKC96] propose an analytical model for such a reversible scheme based on a multidimensional Markov chain. However, the state explosion problem prevents from analyzing large systems (i.e., macrocells covering a large number of microcells). Simulations were conducted to evaluate the performance of the system. A reversible system can accommodate higher traffic for one percent blocking probability. However the number of handover increases by twenty percent. Furthermore, the behavior of the two systems tend to be similar at high load. The benefits of macrocell-to-microcell handover do not seem interesting compared to the increase in signaling.

  In [JF97], a general reversible is analyzed: two mobility behaviors are considered (high-speed and low-speed terminals) and both micro-to-macro and macro-to-micro handover are allowed. In order to reduce the unnecessary handover rate, the take-back process is only performed when a terminal is entering a new microcell. The take-back process improves the probability of forced termination for fast mobile stations and tends to provide a more fair treatment between fast and slow mobile stations.

2.4 Conclusion

The handover handling is one of the important issues that must be taken into account when studying wireless mobile networks. In this chapter, the handover procedure along with the main prioritization techniques were presented. Moreover, we dealt with the handover in hierarchical networks, discussing the important related topics studied in the literature.

Next-generation wireless networks are evolving toward IP-based networks. One major challenge in establishing such networks is the provision of fast handover. In the next chapter, mobility issues in wireless IP networks will be presented. The various IP mobility protocols will be identified and detailed.

Chapter 3

IP Mobility Issues

3.1 Introduction

Wireless access to the Internet may outstrip all other forms of access in the near future. It is likely that mobile users will expect similar levels of QoS as users in wired networks. Such a vision presents a number of technical challenges for the mobility protocols in terms of performance and scalability.

"Mobile IP" is a powerful protocol which supports Internet mobility. Mobile IP allows an IP host to roam within the networks, while maintaining its current connections and remaining reachable by the rest of the Internet. However, Mobile IP has some limitations when applied to wide-area wireless networks with high mobility users. In these networks, Mobile IP introduces significant network overhead, in terms of increased delay, packet loss and signaling. Thus, Mobile IP needs to be enhanced.

"IP Micro-mobility protocols" are designed in order to limit the disruption to user traffic during handover and to handle frequent handovers across multiple subnetworks. IP micro-mobility protocols complement Mobile IP by providing fast and also seamless handover control.

In a first step we present the problems that an IP Mobility protocol can face. Then, we describe the Mobile IP and the different micro-mobility protocols. In a second step, we make a comparative study between the different micro-mobility protocols, considering mobility management issues.

In this chapter, we refer to the mobile user to as the mobile node (MN). We also refer to any node communicating with a mobile node to as a corresponding node (CN).

3.2 IP Mobility Protocol

An IP mobility protocol must meet two major mobile node requirements:

1. Seamless roaming: when a mobile node goes out of his network and roams across the networks, a seamless Internet communication must be maintained between the mobile node and his correspondents whatever the access network used by the mobile node is.

2. Application transparency: the application must be transparent to the mobile node's mobility. It should not be aware of any modification in the mobile node's IP address. Moreover, the application should not be disrupted while the mobile roams outside his network.

3.2.1 Why Isn't IP Mobility Simple?

The IP mobility is difficult to achieve. In fact, the difficulty comes from the current allocation of IP addresses. It is noteworthy to mention that an IP address is used to:

- Identify a particular end-system: the mobile node's IP address should not change during a session since the domain name system (DNS) should always point to the same IP address.
- Identify a particular TCP session in an IP mobile node: the mobile node's IP address should not change during a TCP session.
- Determine a route to a destination IP mobile node: the IP address is used in order to route the packets toward the destination IP mobile node. When a mobile node is roaming outside its home network, it should change its address in order to receive datagrams sent to it in the visited network.

As a result, a suitable IP mobility protocol must allocate a permanent IP address to a mobile node during a TCP session. At the same time, the IP mobility protocol must assign a new IP address to the mobile node when it enters a new network in order to route the datagrams towards the roaming mobile node.

The Mobile IP protocol is designed in order to resolve the above contradiction by using two IP addresses for a mobile node :

- The Home Address: is a permanent address used to identify uniquely the mobile node on the Internet. The home address makes the mobile node logically appear attached to its "home network". Note that the home network is the network at which the mobile node *seems* reachable, to the rest of the Internet, by virtue of its assigned home address.
- The Care-of Address (CoA): is a temporary address used to route the datagrams sent to the mobile node in the "foreign network". Note that the foreign network is the network to which the mobile node is attached when it is not in its home network, and on which the care-of address is reachable from the rest of the Internet.

3.3 Mobile IP

Mobile IP is the current standard for supporting the nodes mobility across IP domains while maintaining transport level connections. There are two versions of Mobile IP: Mobile IPv4 based on IPv4 [Per96b] and Mobile IPv6 based on IPv6 [JP01].

Mobile IPv4 defines two "mobility agents"[1] : the Home Agent (HA) and the Foreign Agent (FA).

The Home Agent is a node on the home network that effectively causes the mobile node to be reachable at its home address even when the mobile node is not attached to its home network.

The Foreign Agent is a a mobility agent on the foreign network that can assist the mobile node in receiving datagrams delivered to the care-of address.

The Home Agent and the Foreign Agent support tunneling datagrams using IP in IP encapsulation [Per96a].

[1] A mobility agent is a node (typically, a router) that offers support services to mobile nodes.

The design of Mobile IPv6 is based on the experiences gained from the development of Mobile IPv4, and the new features provided by IPv6 such as an increased number of available IP addresses and additional IP autoconfiguration features. There are some differences between Mobile IPv4 and Mobile IPv6.

In Mobile IPv6, there is no longer any need to deploy foreign agents: the mobiles make use of the enhanced features of IPv6 to operate in any location away from the home network without any support required from the foreign agents.

Mobile IPv6 requires the exchange of additional information. Thus, new IPv6 Destination Options and ICMP messages are defined. The use of IPv6 destination options allows Mobile IPv6 control traffic to be piggy-backed on any existing IPv6 packet, whereas in Mobile IPv4, separate UDP packets are required for each control message. These destination options are the "Binding Update", the "Binding Acknowledgement", the "Binding Require" and the "Home Address".

Support of what is known in Mobile IPv4 as "Route optimization" (see subsection 3.3.1) is built as a fundamental part of Mobile IPv6, rather than being optional as in Mobile IPv4.

3.3.1 Mobile IPv6 Operation

3.3.2 Home Agent Registration

A mobile node is always reachable by its home address, whether it is currently attached to its home link or is away from home. When a mobile node is at home, packets addressed to its home address are routed to it using conventional Internet routing mechanism. When a mobile node is attached to a foreign network, it is addressable by its care-of address in addition to its home address. The subnet prefix of a mobile node's care-of address is the subnet prefix of the foreign network being visited by the mobile node.

The association of the home address of a MN with a CoA of that MN, along with the remaining lifetime of that association, is called "binding". In Mobile IPv6, the HA and the mobile nodes maintain a binding cache that stores other nodes bindings.

While away from home, a mobile node acquires its care-of address through *stateless* or *stateful* Address Autoconfiguration mechanism. The stateless mechanism [TN98] allows a mobile node to generate its own addresses using a combination of the subnet prefix and an "interface identifier". The subnet prefix is advertised by the routers, whereas the interface identifier uniquely identifies an interface on a subnet.

In the stateful address autoconfiguration mechanism, the mobile nodes obtain interface addresses and/or configuration information and parameters from a server. Servers maintain a database that keeps track of which addresses have been assigned to which mobile nodes. The Dynamic Host Configuration Protocol for IPv6 (DHCPv6) is an example of a stateful address autoconfiguration mechanism [DBV+02].

The mobile node must register its care-of address with its Home Agent and with its CN. Thus, the mobile node sends a packet containing a "Binding Update" destination option to the HA and to the CN; the Home Agent (respectively the CN) replies to the mobile node by returning a packet containing a "Binding Acknowledgement" destination option.

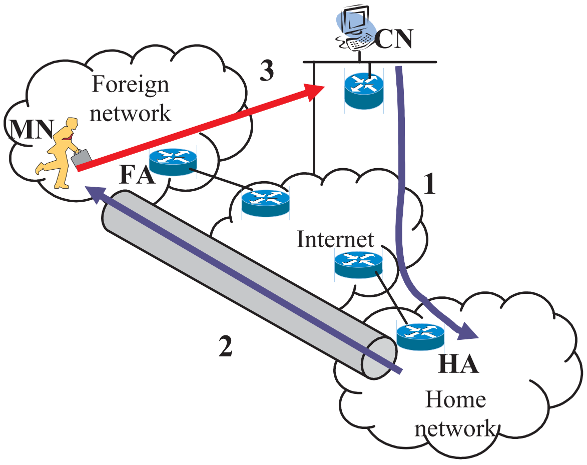
3.3.3 Triangle Routing

The Home Agent must be capable of intercepting any packet addressed to the mobile node's home address. When the HA intercepts packets addressed to the mobile node's home address, it *tunnels* the packets to the mobile node's care-of address. To tunnel each intercepted packet, the home agent encapsulates the packet using IPv6 encapsulation [CD98]. Thus, the home agent, which is the *tunnel source*, inserts a new IP header, or *tunnel header*, in front of the IP header of any packet addressed to the mobile node's home address. The new tunnel header uses the mobile node's care-of address as the destination IP address, or *tunnel destination*. The tunnel source IP address is the home agent address, and the tunnel header uses 4 as the higher level protocol number indicating that the next protocol is again an IP header.

If the mobile node sends packets to the CN, packets are routed directly to the CN address. On the other hand, the packets sent by the CN to the mobile node pass through the mobile node's Home Agent before arriving to destination. This routing is called "Triangle Routing". The triangle routing is a problem since it incurs overhead traffic due to the assistance of the HA to deliver packets to the MN (figure 3.1).

3.3.4 Route Optimization

To avoid the triangle routing, route optimization is applied in Mobile IPv6. When sending a packet to a mobile node, a CN checks its bindings cache for any entry containing the destination address. If a cached binding for this destination address is found, the CN uses an "IPv6 Routing Header" [DH95] (instead of IPv6 encapsulation) to route the packet to the destination address. The use of a Routing Header requires less additional header bytes to be added to the packet, reducing the overhead of Mobile IP packet delivery. The CN sets the fields in the packet's IPv6 header and Routing header as follows:

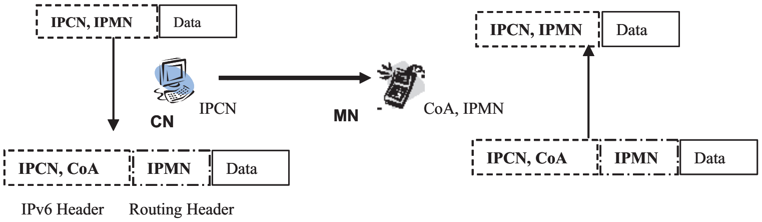


1- CN sends packets to the MN's Home address
2- HA intercepts the packets and tunnels them to the MN's CoA
3- MN sends packets directly to the CN's address

Figure 3.1: Triangle Routing in Mobile IPv4

1. The destination address in the packet's IPv6 header is set to the mobile node's care-of address.

2. The Routing header is initialized to contain a single route segment, with an address of the mobile node's home address.



IPMN is the MN's Home Address

IPCN is the CN 's Home Address

CoA is the MN's care-of address

Figure 3.2: Mobile IPv6 Processing Packets from CN to MN

This packet will be routed to the MN's care-of address, where it will be delivered to the mobile node. Processing the Routing header by the mobile node will then proceed as follows (figure 3.2):

1. The mobile node swaps the destination address in the packet's IPv6 header and the address specified in the Routing header. This results in the packet's IP destination address being set to the MN's home address.

2. The packet is then further processed by the protocol layers and the applications above Mobile IP within the mobile node, in the same way as if the mobile node was at home. *Thus, Mobile IP makes mobility transparent to applications.*

If the CN has no cached bindings for the destination address, the CN sends the packet normally with no Routing Header. The packet is subsequently intercepted and tunneled by the mobile node's home agent as described above. In this case, the mobile node must send a Binding Update to the original sender of the packet. The corresponding node updates then its binding for the mobile node. The subsequent packets sent by the CN will be addressed to the MN's care-of address.

The packets sent by the mobile node are routed directly to the corresponding node. The mobile node sets the source address of the packets to its care-of address and includes the " Home Address" destination option that contains the MN's home address. When the CN receives the packet containing the Home Address option, it exchanges the Home Address field from the Home Address option into the IPv6 header, replacing the original value of the source address field (figure 3.3). These modifications are carried out in order to hide the use of the CoA and of the home address option to further packet processing (e.g. at the transport layer).

3.3.5 Mobile IP Handover

A handover occurs when a mobile node moves from one domain to another. When the mobile node detects that it has moved to a new domain, it obtains a new care-of address in the new domain. Then, the MN informs the home agent about the new CoA through registration. Packets

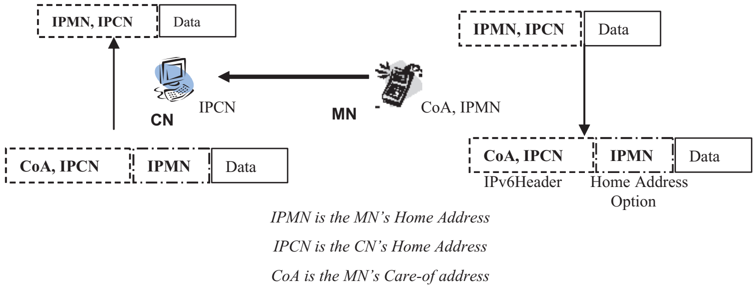


Figure 3.3: Mobile IPv6 Processing Packets from MN to CN

sent to the mobile node are then sent to the new address. Every handover causes a change of the mobile's care-of address. This can be undesirable when the mobile node has a high mobility.

3.3.6 Mobile IP Problems

Mobile IP allows users to roam outside their home networks without disruption to the user's applications. However, when applied to wide-area wireless networks with high mobility users, Mobile IP presents the following problems:

- Latency and control traffic. In Mobile IP, the basic mobility management procedure is the registration to the HA each time the mobile changes network. This process can take a very long time. In the case of fast moving mobiles, the registration process implies a heavy load of control traffic and introduces additional delays. Thus, it is difficult for Mobile IP to support fast handover.

- Quality of Service. The frequent changes of point of attachment and of CoA make it difficult to support QoS for mobile users. With the ReSerVation Protocol (RSVP), for instance, the reservations must be done each time the mobile node changes its CoA, along the entire path. The reservations along the path are done even if the largest part of this path is unchanged. This process implies a heavy load in control traffic and introduces additional delays incompatible with the support of QoS.

- Paging. The paging facilitates efficient power management at the mobile node. In fact, paging allows the mobile node to update the network location less frequently at the cost of providing the network approximate location information. In Mobile IP, the mobile node is expected to update the network on every move. This results in excessive battery power consumption, which is unacceptable for wide-area wireless devices.

To face the above-mentioned problems in IP-based mobile networks, the "Hierarchical IP mobility management" is proposed as precised in the next section.

3.4 Hierarchical Mobility Management

The hierarchical mobility management separates local and wide area mobility in order to improve the performance of mobility management. The hierarchical mobility management defines the "macro-mobility" and the "micro-mobility".

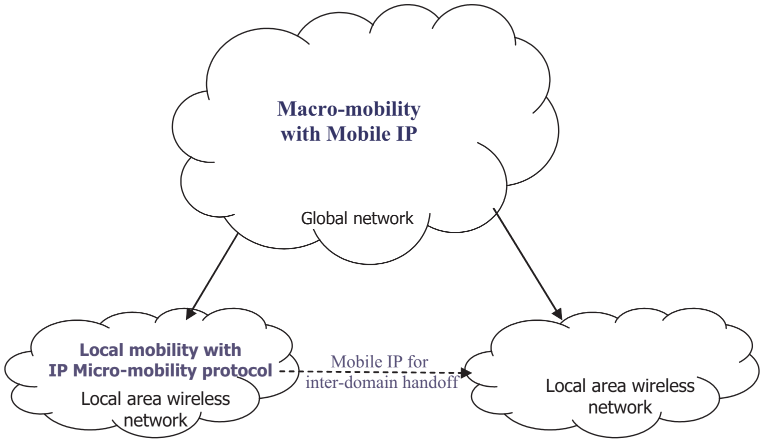

Macro-mobility with Mobile IP
Global network
Local mobility with IP Micro-mobility protocol
Local area wireless network
Mobile IP for inter-domain handoff
Local area wireless network

Figure 3.4: Hierarchical Structure of Macro-mobility and Micro-mobility

Macro-mobility is the mobility on a large scale (e.g. between LANs). Micro-mobility is the local mobility within a single domain. With the hierarchical mobility management, Mobile IP is used to handle the macro-mobility whereas IP micro-mobility protocols handle the micro-mobility (figure 3.4).

The IP micro-mobility protocols handle local mobility *locally* and hide it from home agents. When a mobile node moves from one access point to another one (which is reachable through the same gateway) then the home agent is not informed of the mobile handover. As a result, the local mobility is transparent to the home agent. This considerably reduces the overhead and majority of handovers can be handled locally and faster by the micro-mobility protocols.

The micro-mobility protocols ensure that packets arriving at the gateway are routed to the actual mobile node's point of attachment. Each node in the micro-mobility network maintains a list of node entries. Each entry contains a pointer to the next node toward the mobile node's actual point of attachment. To forward a downlink packet, nodes must read the original destination address in this packet, find the corresponding entry and forward the packet to the next node.

3.5 Micro-mobility Protocols

There are many protocols proposed for micro-mobility in the literature. In this section, we detail five well-known IP micro-mobility protocols: "Cellular IPv6", "HAWAII", "Hierarchical Mobile IP (HMIP)", "Fast Handovers for Mobile IPv6" and "HMIPv6".

3.5.1 Cellular IPv6

Cellular IPv6 [SGCW00] is a micro-mobility protocol relying on Mobile IPv6 for the macro-mobility management. Cellular IPv6 inherits cellular technology principles for mobility management, passive connectivity which permits the mobiles in idle state to be reachable for incoming packets and handover support, but implements these around the IP paradigm.

3.5.1.1 Network Architecture

A Cellular IP network (figure 3.5) comprises a gateway router that connects the Cellular IP network to the Internet as well as several Cellular IP nodes that are responsible for the Cellular IP routing and mobile nodes which support the Cellular IP protocol. A Cellular IP node that also has a wireless access point is called "Base Station". Each Cellular IP node has an uplink neighbor to which it relays the packets originating from the mobile nodes and one or more downlink neighbors to which it relays the packets destined for a mobile node.

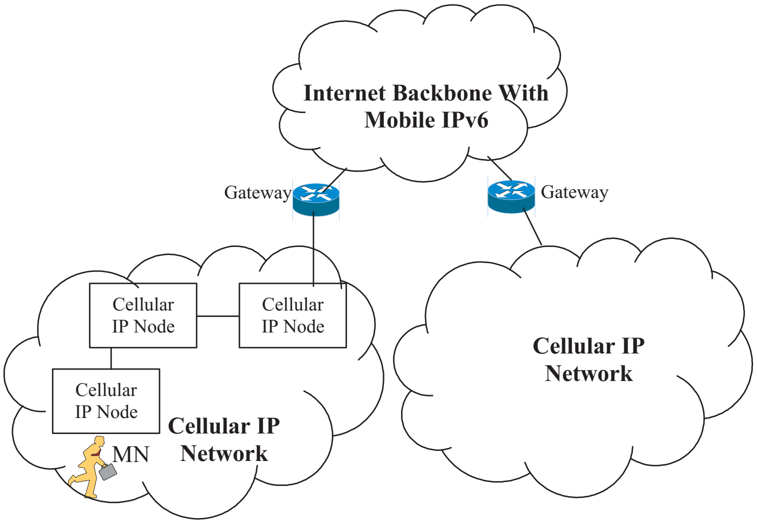


Figure 3.5: Cellular IPv6 Networks

Cellular IP provides two parallel cache systems that store the host-based routing entries. These caches are the *Route cache* and the *Paging cache*. Both types of caches consist of 5-tuples, called "mappings" { IPv6 address, interface, MAC address, expiration time, timestamp }. The IPv6 address is the address of the mobile node to which the mapping corresponds. The interface and the MAC address denote the downlink neighbor toward the mobile node. The timestamp field contains the timestamp of the control packet that has established the mapping.

Each Cellular IP node has a Routing cache, whereas few nodes maintain Paging caches.

While connected to a Cellular IP network, a mobile node must be in one of two states: "active" or "idle". The mobile node moves from idle to active state when it receives or wishes to send a data packet. Active state is maintained as long as the mobile is transmitting or receiving data packets. When the mobile node has not received or transmitted any data packets for some time (the value of this time is implementation-specific), then it returns to the idle state.

3.5.1.2 Location Management Routing and Paging

When an active mobile node wants to *update* the Cellular IP network's Route caches, it has to set up a routing path from the gateway router to its current attachment point. This is done in a reverse manner by sending a *route-update message* from the mobile node to the gateway router. The route-update message is received by the base station. It is then forwarded in a hop-by-hop manner following the uplink neighbors of each Cellular IP node towards the gateway router.

Whenever a route-update message passes through a Cellular IP node, a mapping for the related mobile node is written in the Paging and Route caches. When the route-update message arrives at the gateway router, the gateway router adds a mapping to its caches and drops the message afterwards. Nodes in a Cellular IP access network maintain a distributed hop-by-hop location database that is used to route packets to mobile nodes.

The mappings in the Cellular IP caches are soft state. This means that after a certain expiration time they are not valid any more. This is necessary since due to a link loss a mobile node might not be able to tear down its mappings before leaving the network. In order not to lose its routing path, a mobile node has to *refresh* its mappings periodically. Thus, the active node sends regular data packets or sends *periodic* route-update message if the active mobile node has no data to transmit. The data packets and the route-update message refresh the mapping in the Route and Paging caches by setting the expiration time to the sum of the current time and the caches timeout. The caches timeout is the validity time of mappings in the caches. Note that the Paging caches have longer timeout than the Route caches.

Idle mobile nodes send *periodic paging-update* messages that refresh the Paging caches. Note that Paging caches are updated by all uplink update packets (route-update and paging-update) and refreshed by all uplink packets including data packets as well (table 3.1). It is to be noted that route-update-time (respectively paging-update-time) is the time between consecutive route-update packets (respectively paging-update packets).

In Cellular IP nodes where both a Route and a Paging cache are maintained, the Route cache mappings are used to route the packets in the downlink direction.

If a packet addressed to a mobile node arrives at a Cellular IP node, then the node proceeds as follows:

- If no up-to-date Route cache mapping is available for the mobile node, the Paging cache is then used to route the packet. This is called an “implicit paging”.

- If the node does not maintain a Paging cache and does not have up-to-date Route cache mappings, the node then broadcasts the packet to all its downlink neighbors. By this mechanism, groups of several base stations are built in which idle mobile nodes are searched when a packet has to be delivered to them. A group of base stations are called a paging area. To minimize paging traffic, an idle mobile node has to update its Route cache entries immediately after receiving a paging packet.

When a mobile node is in an active state, the Cellular IP location management has to follow its movement from one base station to another to be able to deliver packets without searching for the mobile node. As a consequence, active mobile nodes must notify the network about each handover. Thus, mobile nodes must transmit a route-update packet when they cross cell boundaries. For idle mobile nodes, exact location tracking is less important. Instead minimizing communication to save battery power has higher priority. The idle mobile nodes must transmit a paging-update message when they move to a new paging area.

3.5.1.3 Cellular IP Handover

A handover occurs when there is a change of access point during an active transmission. Cellular IP supports two types of handover schemes.

	Paging Caches	**Route Caches**
Refreshed by	all uplink packets (data, paging-update, route-update)	data and route-update packets
Updated by	all update packets (paging-update, route-update)	route-update
Updated when	moving to a new Paging area or after paging-update-time	moving to a new cell or after route-update-time
Scope	idle and active mobile nodes	active mobile nodes
Purpose	route downlink packets if there is no Route cache entry	route downlink packets

Table 3.1: Route and Paging caches

1. "Cellular IP hard handoff" uses an algorithm that trades off some packet loss in exchange for minimizing handover signaling.

   When a mobile node switches to a new base station, it sends a route-update packet to make the chain of cache bindings point to the new base station. Packets that are traveling on the old path will be delivered to the old base station and will be lost. Although this loss may be small, it can potentially degrade the TCP and UDP throughput. This degradation is severe for TCP since a data loss incurs data retransmission.

2. "Cellular IP semi-soft handoff" is proposed for the wireless technologies that provide simultaneous connections. Semi-soft handoff tries to pro-actively notify the new access point before actual handover. Semi-soft handoff minimizes packet loss and provides improved TCP and UDP performance. During semi-soft handoff a mobile node may be in contact with either the old or the new base station and receive packets from them. Packets intended to the mobile node are sent to both base stations. When the mobile node eventually moves to the new location, it can continue to receive packets without interruption.

   To initiate semi-soft handoff, the moving mobile node transmits a route-update packet to the new base station and continues to listen to the current one. A flag, denoted by S, is set in this route-update packet to indicate semi-soft handoff. Semi-soft route-update packets create new mappings in the Route and Paging Cache similarly to regular route-update packets. When the semi-soft route-update packet reaches the cross-over node where the old and new path meet, the new mapping is added to the cache instead of replacing the old one. If the path to the new base station is longer than to the current base station or it takes a non negligible time to switch to the new base station, then some packets may not reach the mobile node. To overcome the problem, packets sent to the new base station can be delayed during the semi-soft handoff. This way, a few packets may be delivered twice to the mobile nodes, but in many cases this results in better performance than in the case where few packets are lost.

When the mobile node eventually makes the move, then the packets will already be underway to the new base station and the handover can be performed with minimal packet loss. After migration, the mobile node sends a route-update packet to the new base station with the S bit cleared. This route-update packet will remove all mappings in Route Cache except for the ones pointing to the new base station. The semi-soft handoff is then complete.

For the wireless technologies that do not provide the simultaneous connections, the "indirect semi-soft handoff" is proposed. When the mobile decides to make a handover, it sends the packet to the current BS instead of sending it directly to the new BS. This packet will have as a destination IP address, the IP address of the new BS. A flag denoted by I is set in the route-update packet to indicate indirect semi-soft handoff. The current BS will forward this packet uplink to the Gateway normally. The Gateway then uses normal IP routing to deliver the packet to the new BS. When the new BS receives the indirect handover packet, a semi-soft route-update packet is created with the IP address of the mobile node. It is then forwarded upstream. The algorithm then proceeds as in the semi-soft handoff algorithm.

3.5.2 Handoff-Aware Wireless Access Internet Infrastructure (HAWAII)

3.5.2.1 Network Architecture

HAWAII from Lucent Technologies is a micro-mobility protocol that relies on Mobile IPv4 to provide wide-area inter-domain mobility ([RLPTV99], [RLPST$^+$99]).

In HAWAII a hierarchy based on domains is used as depicted in figure 3.6. The gateway in each domain is called "Domain Root Router" (DRR). A HAWAII domain comprises several routers and base stations running the HAWAII protocol, as well as mobile nodes.

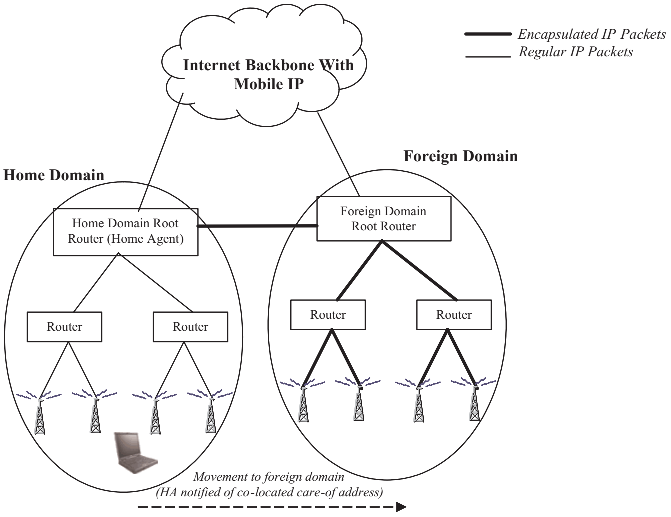

Internet Backbone With Mobile IP
Encapsulated IP Packets
Regular IP Packets
Home Domain
Foreign Domain
Home Domain Root Router (Home Agent)
Foreign Domain Root Router
Router
Router
Router
Router
Movement to foreign domain (HA notified of co-located care-of address)

Figure 3.6: HAWAII Network Architecture

3.5.2.2 HAWAII Location Management and Routing

Upon entering a new foreign domain, the mobile node is assigned a “co-located care-of address”. Note that a co-located address is a care-of address acquired by the mobile node as a local IP address, which the mobile node associates with one of its own interfaces. When using a co-located care-of address, the mobile node serves as the endpoint of the tunnel and itself decapsulates the packets tunneled to it. The mode of using a co-located care-of address allows a MN to function without a foreign agent. However, it places additional burden on the IPv4 address space because it requires a pool of addresses within the foreign network to be available to visiting mobile nodes.

The care-of address of the mobile node remains unchanged while the mobile node is moving within the foreign domain. “Specialized path setup schemes” are employed to establish and update host-based routing entries for the mobile nodes in the routing tables of the routers.

There are three types of HAWAII path setup messages: powerup, update and refresh. On power up, a mobile node sends a *path setup powerup* message to the domain root router which is processed in a hop-by-hop manner. This powerup message adds a routing entry for the concerned mobile node on all routers on its way to the domain root router. The packets arriving at the domain root router based on the subnet address of the domain are then routed using the host-based routing entries established.

The router entries are soft state, i.e. they have to be refreshed periodically by *path setup refresh* messages. The mobile node sends periodic path setup refresh messages to the base station to which it is attached. The base stations and the intermediate routers, in turn, send periodic aggregate refresh in a hop-by-hop manner towards the domain root router .

HAWAII uses *path setup update* messages to establish and update host-based routing entries for the mobile nodes in selected routers in the domain, so that packets arriving at the domain root router can reach the mobile node with limited disruption. The choice of when, how, and which routers are updated constitutes a particular path setup scheme. In subsection 3.5.2.4, we describe four path setup update schemes.

Unlike Cellular IP, HAWAII does not replace IP but works above IP. Each node inside the network must not only act as a classical IP router but also maintains mobility routing information (see subsection 3.5.2.4). In this sense, HAWAII nodes can be considered as enhanced IP routers.

3.5.2.3 HAWAII Paging

With HAWAII, mobile nodes can switch to a *standby state*. Mobile nodes in standby state have to notify the network of a paging area change and not of each base station handover. In this case, the network does not have to keep exact location information of those mobile nodes, but only information about their approximate location. When a packet arrives for a mobile node in a standby state, the network has to page the mobile node before it delivers the packet. This paging induces the mobile node to switch to the active state immediately. For using HAWAII’s paging support, it is necessary to have link-layer paging functionality on the wireless link which means that the mobile node is able to identify its paging area and to detect paging requests.

A typical solution for identifying the paging area is, that base stations periodically send beacon signals including the paging area identities on a broadcast channel. Thus, a mobile node listening to this channel can easily detect a change. The paging requests of the base stations can be sent on separate paging channels to which the mobile nodes are listening.

The network has to maintain paging information for each mobile node and has to deliver paging requests for these nodes to the base stations. One way to achieve this is to deliver the paging requests to each base station within the area using a unicast message to each one. Because that would be a waste of bandwidth, HAWAII relies on the IP multicast routing protocol. Each paging area is assigned a multicast group address. All base stations within that paging area join this multicast group.

Some network nodes maintain a Paging cache that maps the mobile node's IP address to the IP multicast group. When a packet sent to a mobile node arrives at a router that does not have an active route entry and that has an active paging entry, the router then detects then that the MN is in standby state. If the packet arrives from DRR and if the router is a Base Station or is a part of a multicast tree with more than two branches, the paging is then initiated: the router buffers the packet and multicasts a paging message to awake the mobile.

3.5.2.4 HAWAII Handover

A handover occurs when the mobile node's next hop IP node changes. Before describing re-establishment of routes after intra-domain movements, let us define the *crossover* router. In HAWAII, the crossover router is the closest router to the mobile node, that is at the intersection of two paths: one path is from the domain root router to the old base station and the second path is from the old base station to the new base station.

HAWAII defines four alternative path setup schemes that control handover between access points. An appropriate path setup scheme is selected depending on the operator's priorities that are eliminating path losses, minimizing handover latency and maintaining packet ordering. The four path setup schemes considered can be classified into two types of schemes, *forwarding* and *nonforwarding* schemes. These schemes are based on the way packets are delivered to the mobile node during a handover.

In the forwarding schemes, packets are forwarded from the old base station to the new, whereas in the nonforwarding schemes, they are diverted at the crossover router to the new base station. There are two variants of the forwarding schemes: the Multiple Streams Forwarding (MSF) and the Single Stream Forwarding (SSF).

With the MSF scheme and during handover, the new base station sends a path setup update message directly to the old base station whose address was transmitted by the mobile node (figure 3.7). The old base station performs a table look-up for a route to the new base station and determines the next hop router. It adds a routing table entry for the mobile node pointing to that next hop router and forwards the path setup update message.

From this stage on, the old base station forwards all data packets for the concerned mobile node to the new base station according to the new forwarding entry. The next hop router performs similar actions and in that way the packet is forwarded up to the crossover router which changes the moved mobile node's routing entry, too. From this stage on, the crossover router diverts new data packets to the new base station. The crossover router then forwards the path setup update message completely to the new base station that also adds a forwarding entry.

The mobile node continues to receive packets in transit during handover. With this scheme, packets forwarded by the crossover router may arrive before older packets forwarded by the base station. This scheme leads to the creation of multiple misordered streams during handover.

The SSF scheme updates the routing entries such that packets are forwarded from the old base station to the new base station in a single stream. In order to achieve this, the authors in [RLPST$^+$99] use a technique called "interface-based forwarding". This technique requires more descriptive routing table entries. A routing table typically has an entry of the form *(IP address, outgoing interface)*. In this scheme, the router must be able to route based on an additional field, the incoming interface of the packet. The resulting routing entry is of the form *(incoming interface(s), IP address, outgoing interface)*.

In Figure 3.8, messages 1-5 establish these entries resulting in packets arriving at the old base station and being forwarded to the new base station as a single stream. The old base station subsequently sends message 6 to Router 0 for diverting the stream at the crossover router. While this scheme is also lossless and maintains a single stream of forwarded packets until the diversion is performed at the crossover router (until message 6), it is somewhat complex to implement.

In the nonforwarding schemes, as the path setup message travels from the new base station to the old base station, data packets are diverted at the crossover router to the new base station, resulting in non forwarding of packets from the old base station. There are two variants of the nonforwarding scheme, motivated by two types of wireless networks. The unicast nonforwarding (UNF) scheme is optimized for networks where the mobile node is able to listen/transmit to two or more base stations simultaneously for a short duration (e.g. code division multiple access (CDMA) network). The multicast nonforwarding (MNF) scheme is optimized for networks where the mobile node is able to listen/transmit to only one base station as in the case of a time-division multiple access (TDMA) network.

The UNF scheme is illustrated in Fig. 3.9. In this case, when the new Base Station receives the path setup message, it adds a forwarding entry for the mobile node. This entry contains the mobile node's IP address and the interface on which the BS received this message. The BS then performs a routing table lookup for the old base station and determines the next hop router, Router 2. The new base station then forwards message 2 to Router 2. This router performs similar actions and forwards message 3 to Router 0. At Router 0, the crossover router in this case, forwarding entries are added such that new packets are diverted directly to the mobile node at the new base station. Eventually, message 5 reaches the old base station which then changes its forwarding entry and sends an acknowledgment, message 6, back to the mobile node.

The second non forwarding scheme MNF is very similar to the UNF scheme. The main difference is that the crossover router, Router 0, multicasts data packets for a short duration. In Fig. 3.10, Router 0 dualcasts data packets from interface A to both, the new and old base stations after it receives message 3 and until it receives message 6. This helps in limiting packet loss in networks in which the mobile node can only listen to a single base station at a time.

3.5.3 Hierarchical Mobile IP (HMIP)

The Hierarchical Mobile IP (HMIP) proposal from Ericsson and Nokia employs an hierarchy of foreign agents to locally handle Mobile IPv4 registration [GJP02]. Figure 3.11 assumes two

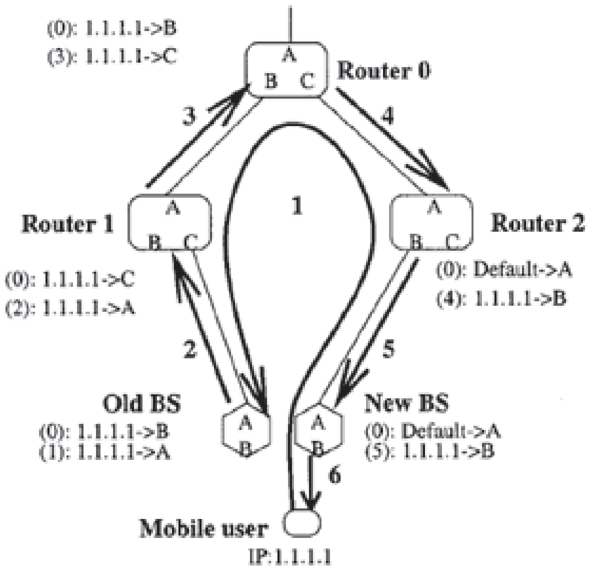


Figure 3.7: Multiple Stream Forwarding Scheme

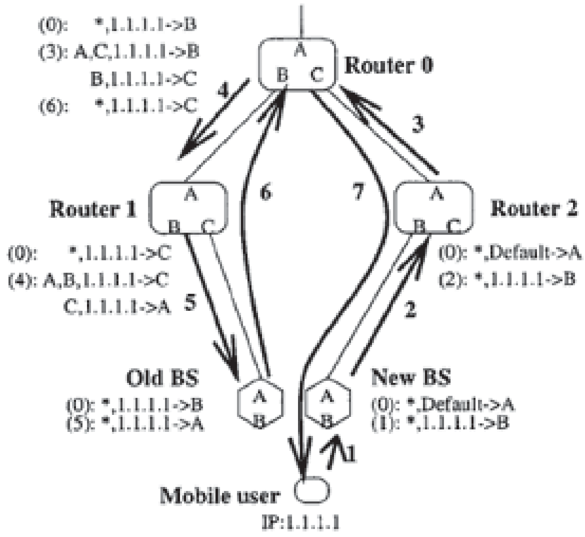


Figure 3.8: Single Stream Forwarding Scheme

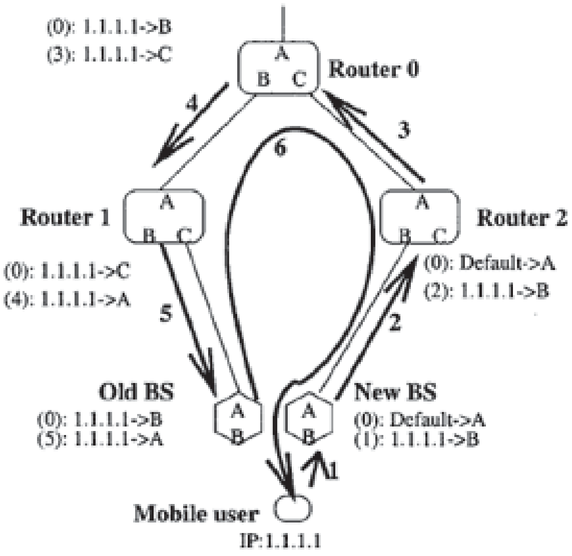


Figure 3.9: Unicast Nonforwarding Scheme

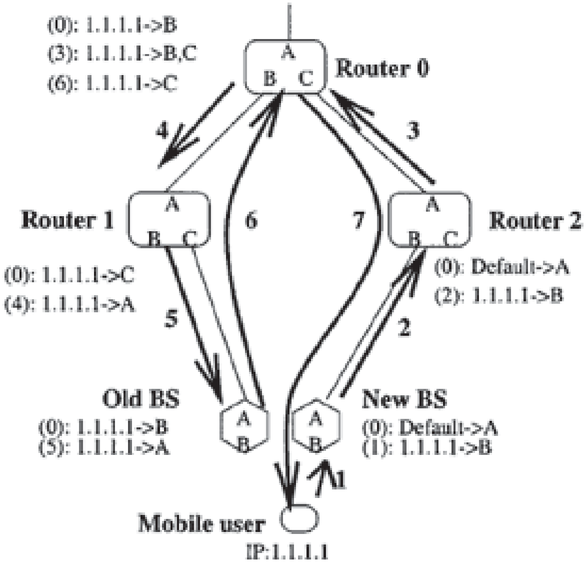


Figure 3.10: Multicast Nonforwarding Scheme

hierarchy of foreign agents in the visited domain: the HMIP network consists of several foreign agents connected to a new network entity called a “Gateway Foreign Agent” (GFA).

The principle of HMIP is that the MN that connects for the first time to the domain registers to its HA with the address of the GFA as CoA. This is a *Home Registration*. Since the care-of address registered at the home agent is the GFA address, this CoA will not change when the mobile node changes foreign agent under the same GFA. Thus, the home agent does not need to be informed of any further mobile movements within the visited domain.

Inside the foreign domain, the MN will only perform *Regional Registrations*. The mobile sends a regional registration to the GFA each time the MN changes a FA. The registration contains the new local CoA of the MN. This address can be either a co-located address or the FA’s address. This address is used by the GFA to join the MN while MN remains connected to the same FA. The routing with Hierarchical Mobile IP is then very simple. The packets addressed to the MN are first intercepted by the HA and tunneled to the GFA. The GFA decapsulates and tunnels these packets towards the current local CoA of the MN.

Hierarchical Mobile IP also includes several levels of hierarchy of FAs between the leaf FA and the GFA (figure 3.12). The mobile node sends mobile IP registration messages with appropriate message extensions to update its location information. These registration messages establish tunnels between foreign agents which are on the path from the mobile node to the gateway foreign agent.

When a corresponding node sends traffic to the mobile node, the traffic arrives at the HA, and then the HA tunnels the traffic to the GFA. The GFA or the FA at each level of the hierarchy has a “visitor list” for the mobile node. The “visitor list” contains among others the MN’s home address, the IP destination address, the remaining registration lifetime and the address of the next lower foreign agent in the hierarchy of FAs. The datagram arriving at a GFA will be re-routed at the next lower level of the hierarchy: the datagram will be decapsulated and then re-encapsulated by the GFA. This re-routing occurs at each level of the hierarchy until the datagram reaches the last point which is either the MN itself (in case of a co-located CoA) or a FA that can deliver the datagram to the MN with no further special Mobile IP handling.

3.5.4 Hierarchical Mobile IPv6 (HMIPv6)

Hierarchical Mobile IPv6 is proposed by Ericsson and INRIA [SCEMB01]. In Mobile IPv6 there are no Foreign Agents, but there is a need to provide a local entity to assist Mobile IP handovers and to reduce the mobility signaling (figure 3.13). With HMIPv6, a new Mobile IPv6 node, called the “Mobility Anchor Point” (MAP), is used and can be located at any level in a hierarchical network of routers, including the Access Routers (AR) level.

Upon arrival to the foreign network, the MN enters a “MAP discovery phase”. During this phase, the MN receives messages called “Router Advertisements” communicated by the access routers. These Router Advertisements will help the MN discover the global address of the MAP, the MAP’s subnet prefix, the distance of the MAP from the MN and the preference for this particular MAP. Note that the preference for a MAP is set based on local policies such as node overload. A MAP with a preference value of 15 should not be used.

A mobile node has two care-of addresses in a Hierarchical Mobile IPv6 network. These CoA addresses are the Regional Care-of address (RCoA) and the Link Care-of address (LCoA). These

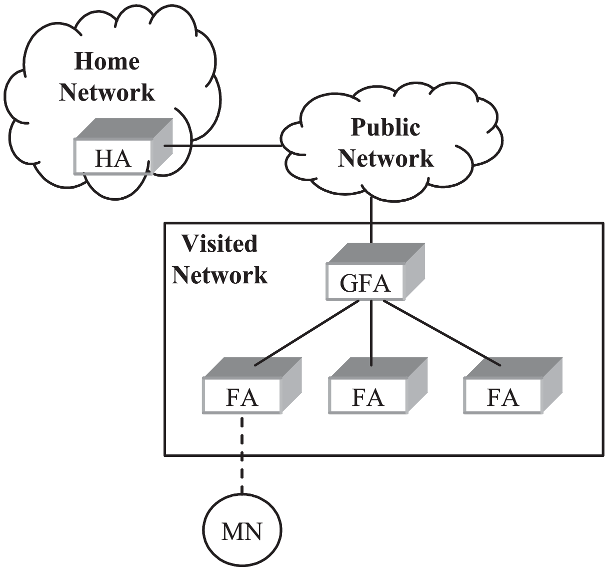


Figure 3.11: HMIP with one level of FA hierarchy

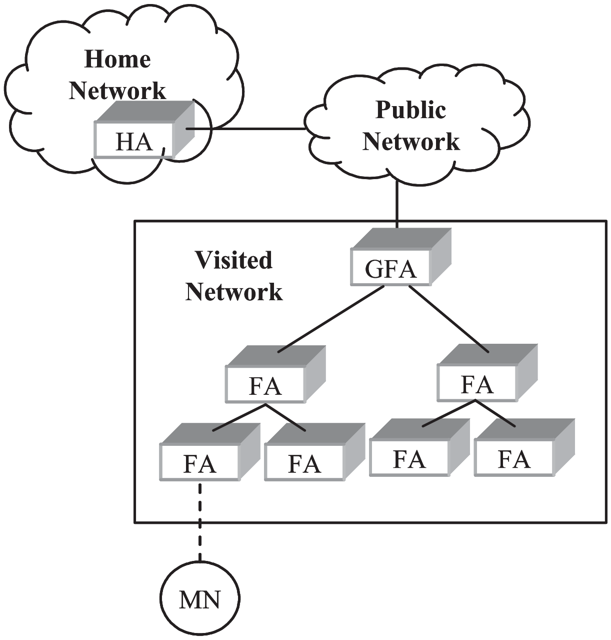


Figure 3.12: HMIP with multi-level FA hierarchy

CoAs are configured as follows:

- The LCoA is configured on a MN's interface based on the prefix advertised by its access router.
- Two different MAP modes are proposed based on the usage of the RCoA. These modes are the Extended mode and the Basic mode. In the Extended mode, the RCoA is assigned to one of the MAP's interfaces. In the Basic mode, the RCoA is auto-configured by the MN.

3.5.4.1 Basic Mode

In the basic mode, the MN has two addresses: the RCoA and the LCoA. The RCoA is formed in a stateless manner by combining the MAP's subnet prefix to the MN's interface identifier. After forming the RCoA, the MN sends a Binding Update (BU) to the MAP. The BU specifies the binding between the RCoA and the LCoA. The MAP will then perform the "Duplicate Address Detection" [2](DAD) for the MN's RCoA on its subnet. If successful, the MAP will return a Binding Acknowledgment (BA) to the MN indicating a successful registration.

The MN must register its new CoA with its HA and its corresponding nodes by sending a BU that specifies the binding of the MN's Home Address with the MN's RCoA.

The MAP acts exactly like a HA. It then intercepts all packets addressed to the RCoA of the registered mobile nodes. The MAP encapsulates the intercepted packets and routes them to the corresponding LCoA. The MN will then decapsulate the packets.

[2] Duplicate Address Detection (DAD) verifies the uniqueness of an unicast address prior to assigning it to an interface [TN98].

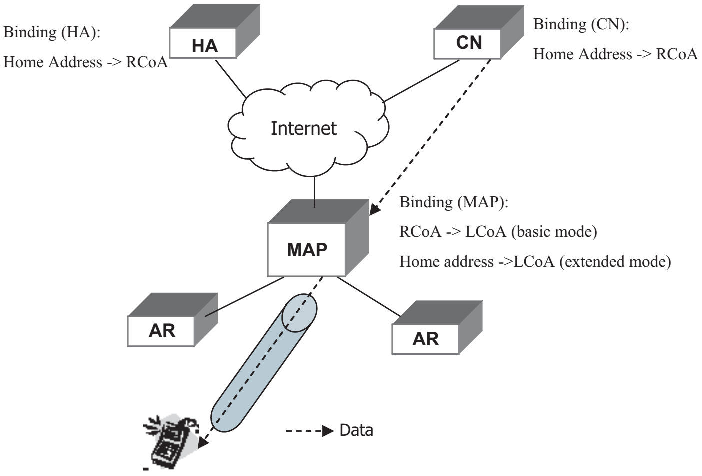


Figure 3.13: Hierarchical Mobile IPv6 Network

3.5.4.2 Extended Mode

In this mode, the MN has a RCoA assigned to one of the MAP's interfaces. Hence, the MN must not use its RCoA as a source address in its outgoing packets.

After the MAP discovery phase, the MN sends a BU to the MAP. The BU specifies the binding of the MN's home address with the MN's LCoA.

After receiving the BA from the MAP, the MN sends a BU to the HA. This BU specifies the binding of the MN's home address with the MN's RCoA.

All packets directed to the MN will be received by the MAP and tunneled to the MN.

If the MAP receives packets containing a routing header from a CN, the MAP processes the routing header as specified in [DH95] and encapsulates the packet to the MN. Upon reception of an encapsulated packet with no routing header from a MN's HA, the MAP decapsulates the packet. Then, the MAP tunnels the packet to the MN's registered LCoA if the inside packet contains a destination address that belongs to the MAP.

3.5.4.3 Hierarchical Mobile IP handover

Fast Handovers can be supported with HMIPv6. Fast Handovers are required to ensure that the layer 3 handover delay is minimized. Fast Handovers mimimize and possibly eliminate the period of service disruption which normally occurs when a MN moves between two access routers. This period of service disruption usually occurs due to the time required by the MN to update its HA using Binding Updates after it moves between access routers. The mechanism to achieve Fast Handovers is explained in the next section.

3.5.5 Fast Handovers for Mobile IPv6

Fast Handovers for Mobile IPv6 is proposed by Ericsson, Nokia, Sun Microsystems, and Cisco Systems [DYP$^+$02]. The Fast Handovers for Mobile IPv6 introduces some terminology used throughout this subsection.

AR refers to an access router that is the last router between the network and the mobile node, i.e, the MN has link layer connectivity to the access router.

oAR refers to an old Access Router, i.e, the AR involved in handling a MN's traffic prior to a Layer 2 (L2) handover.

nAR refers to a new Access Router, i.e, the AR anticipated to be handling a MN's traffic after completion of an L2 handover.

oCoA refers to an old Care of Address, i.e, the care-of Address prior to the MN's first movement. During fast handover, it may be reused until the MN determines an appropriate time to change it, even if the MN changes subnet.

nCoA refers to a new Care of Address, i.e, the care-of Address in the new subnet. During fast handovers, configuration with a nCoA may be delayed while the MN is engaged in latency-sensitive real time traffic or is rapidly moving across a series of ARs.

RtSolPr refers to Router Solicitation for Proxy. It is sent by the MN to the oAR when the MN has information that is about to be handed over to another AR.

PrRtAdv refers to Proxy Router Advertisement. It is sent by the oAR to the MN, either in response to RtSolPr or as a result of information available to the MN that the MN is about to be handed over to another AR.

During handover, there is a time period during which the mobile node is unable to send or receive IPv6 packets due to the lack of a solid layer 2 connection and to the time required at layer 3 to re-establish the mobile node's care-of address on the new subnet. This time period is referred to as handover latency. In order to minimize the handover latency, two mechanisms are described:

1. Anticipated Handover. With this mechanism, Layer 3 initiates handover to the new access router while the MN still has Layer 2 connectivity to the current access router. Either the mobile node or the current access router has predictive information of the actual Layer 2 handover about where the mobile will be moving. Moreover, the mobile node or the current access router can force handover to a particular new access router.

2. Tunnel-based Handover. With this mechanism, the MN defers Layer 3 handover until it is on the new access router. The current access router tunnels packets to the mobile node under its old CoA. The tunneling mechanism is maintained until the MN performs Layer 3 handover.

3.5.5.1 Anticipated Handover

Anticipated handover can be either network initiated or mobile initiated.

1. In mobile initiated handover, the MN has predictive information about the next point of attachment, or it chooses to force movement to a new point of attachment. The MN initiates signaling to the old Access Router: the MN sends a RtSolPr to the oAR, and receives a PrRtAdv in response, providing the MN with the Layer 3 information, such as the new CoA allocated in the new subnet.

2. In network initiated handover, the oAR has predictive information about the next point of attachment to which the MN will move. Thus, oAR sends an unsolicited PrRtAdv to the MN.

There are two possible ways of handling CoA configuration in the new subnets.

1. Addresses on the new subnet are allocated using the IPv6 Stateless Address Autoconfiguration mechanism. In this case, the oAR must construct a nCoA based on the MN's interface ID and the nAR's subnet prefix.

2. Addresses may be allocated statefully using DHCPv6. In this case, the nAR must return the nCoA.

When the oAR receives an indication from Layer 2 that the MN will be moving (network initiated handover) or a RtSolPr from the MN indicating that the MN wants to move (mobile initiated handover), the oAR exchanges messages with the nAR in order to obtain or validate the new CoA for the MN (figure 3.14). The oAR sends a Handover Initiate (HI) message to the nAR. The HI message contains the requested nCoA on the new subnet, if stateless address configuration is in use, and the oCoA being used at the oAR.

When the nAR receives HI, it does the following:

1. If the HI message does not have a nCoA, it allocates a nCoA.

2. If the HI message contains a proposed nCoA, the nAR validates the nCoA.

The nAR replies to the oAR with a Handover Acknowledgment (HACK) message containing either the nCoA allocated by it, or an indication whether the nCoA proposed by the oAR is valid or not.

The timing of when the oAR sends the PrRtAdv to the MN depends on whether stateless or stateful address configuration is used.

In case of stateful address allocation, the oAR obtains the nCoA from the nAR through HI/HACK message exchange. The HI/HACK messaging must be completed before transmitting the PrRtAdv to the MN.

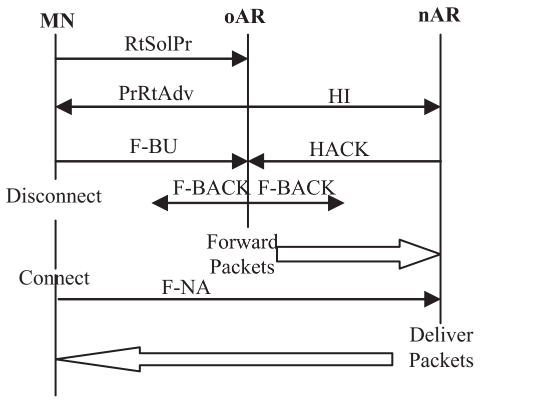
In case of stateless address configuration, the oAR may send the PrRtAdv prior to completing the HI/HACK message exchange.

If the HACK indicates that the nCoA is valid, the oAR must *prepare* to forward packets for the MN to the nCoA. If the HACK indicates that the nCoA is not valid, the oAR must *prepare* to tunnel packets for the MN to the oCoA at nAR.

As soon as the MN receives confirmation of a pending Layer 3 handover through the PrRtAdv and has a nCoA, the MN sends a Fast Binding Update (F-BU) to oAR before the Layer 2 handover is executed.

On receipt and validation of the F-BU, the oAR responds with a Fast Binding Acknowledgment (F-BACK). The oAR waits for a F-BU from the MN before forwarding packets. On receipt of the F-BU, the oAR forms a temporary tunnel for the lifetime specified in the F-BACK, and the F-BACK is sent through the tunnel to the MN on the new link. The F-BACK may also be sent to the MN over its old link, in case the MN has not yet moved.

When the MN arrives on the new subnet and its Layer 2 connection is ready for Layer 3 traffic, it sends a Fast Neighbor Advertisement (F-NA). The nAR may deliver packets to the MN as soon as it receives an indication from Layer 2 that the link is solid enough for it to begin sending packets, or it may start packet delivery when it receives a F-NA from the mobile node.



1- RtSolPr exists in case of mobile initiated handover.
2- The timing of HI/HACK depends on whether stateless or stateful address configuration is in use.

Figure 3.14: Anticipated Handover

3.5.5.2 Tunnel-based Handover

In tunnel-based handover, the mobile moves to its nAR, without the involvement of Layer 3 on the MN or ARs. The MN continues to receive its packets without having to change its CoA. Bi-directional edge tunnels (BET) between the ARs assure that packets are delivered to the MN and sent to the CNs. The MN may delay obtaining a nCoA until its IP traffic, and in particular real-time IP traffic, is reduced or is absent, or until the MN is moving slowly enough that it has time to obtain a new CoA.

This approach allows the MN to move rapidly across a connected series of subnets without any Mobile IP related signaling traffic over the air. Because tunnel-based handover depends on bi-directional edge tunnels, it is called Bi-directional Edge Tunnel Handover (BETH).

Let us present the Layer 2 trigger before describing the progress of a tunnel-based handover mechanism. In fact the L2 trigger is an information from L2 that informs Layer 3 (L3) of the detailed events involved in handover sequencing at L2. L2 triggers are not specific to any particular L2, but rather represent generalizations of L2 information available from a wide variety of L2 protocols.

L2 triggers BETH requires an L2 Source Trigger (L2-ST) at oAR prior to L2 handover start or an L2 Target Trigger (L2-TT) at nAR prior to L2 handover completion. It also requires an L2 Link Down Trigger (L2-LD) at oAR when the old link to the MN is severed, and an L2 Link Up Trigger (L2-LU) on nAR and on MN immediately upon completion of L2 handover. Table 3.2 contains the above-mentioned triggers.

Tunnel-based Handover Progress The progress of a tunnel-based handover is described as follows. Note that the numbered items refer to the steps reported in Figure 3.15.

1. Either the oAR or nAR receives an L2 trigger informing it that a certain MN is about to move. Two cases can be distinguished:

L2 Trigger	Abbreviation
Source Trigger	L2-ST
Target Trigger	L2-TT
Link Down	L2-LD
Link Up	L2-LU

Table 3.2: L2 Triggers

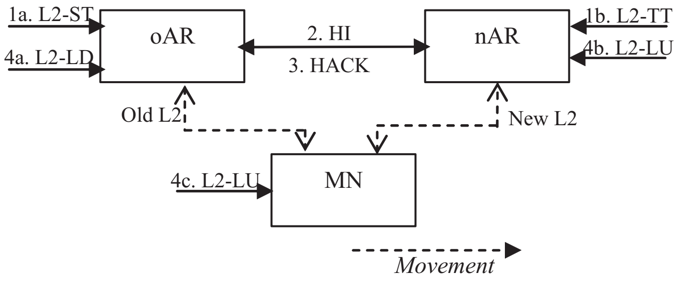


Figure 3.15: Tunnel-based Handover

- 1a- The L2 trigger is a source trigger (L2-ST) at oAR. The trigger contains the MN's L2 address and an IP identifier (L2 address that can be mapped to an IP address or the IP address directly) for nAR.
- 1b- The L2 trigger is a target trigger (L2-TT) at nAR. The trigger contains the MN's L2 address and an IP identifier for oAR.

2. The AR (oAR or nAR) receiving the trigger must send a HI to the other AR. Two cases can be distinguished:

   - 2a- The oAR sends the HI. In this case, the HI includes options that contain the MN's home address, the MN's oCoA, the MN's L2 address and an amount of time. This amount of time is the time the oAR is willing to extend tunnel service to the MN's packets before the nAR renews the bi-directional edge tunnel. If this amount of time is zero, this means that the oAR is not willing to tunnel any packets for MN.
   - 2b- The nAR sends the HI. In this case, the HI includes options that contain MN's L2 address and a request for the amount of time the oAR should extend tunnel service for the MN's packets.

3. The AR (oAR or nAR) receiving the HI must send a HACK to the other AR. There are two cases:

   - 3a- The oAR sends the HACK. The HACK includes options that contain the MN's home address, the MN's oCoA, the MN's L2 address and an amount of time. The latter is the time the oAR is willing to extend tunnel service to the MN's packets before nAR must renew the request. If this amount of time is zero, the oAR is not willing to extend tunnel service to the MN.

- 3b- The nAR sends the HACK within which no information is precised.

4. The start of L2 handover is signaled by an L2-LD trigger at oAR. The completion of L2 handover is signaled by an L2-LU trigger at nAR and MN. Each handles the trigger in the following way:

    - 4a- When the oAR receives the L2-LD trigger, it must begin forwarding MN's packets through the BET to nAR.
    - 4b- When the nAR receives the L2-LU trigger, it must begin delivering packets to the MN and must forward any out-bound packets from MN through the BET to oAR.
    - 4c- Based on its current movement traffic pattern, the MN may either defer obtaining a nCoA or begin the process of obtaining a nCoA.

3.6 Comparison of IP Micro-mobility Protocols

In this section, we propose to compare between the studied IP micro-mobility protocols. We define the major issues for the mobility management that are the basis of our comparison.

- Interaction with Mobile IP. The interaction of the micro-mobility protocols with Mobile IP has a great influence on the overhead introduced in the micro-mobility domain.

  If the Mobile IP tunnel ends at the mobile node, less processing is required by the domain nodes. However, if the tunnel ends at the domain gateway, the latter must decapsulate the packets and re-encapsulate them in order to route them to the mobile nodes. This places additional load on the domain nodes. This is the case of the Hierarchical Mobile IP (HMIP) protocol where multiple re-encapsulations and decapsulations are needed in order to deliver the packets to their destination.

- Protocol layers. The choice of the protocol layers that support the nodes in the micro-mobility domains has important implications related to the network management. The nodes reuse the network management, traffic engineering and quality of service features supported by a particular protocol layer. For instance, nodes operating at L3 can employ differentiated services per hop behaviors. In addition, the choice of protocol layers influences the device availability, the device cost and the type of encapsulation required by the nodes.

- Paging. Notifying the network on each handover regardless of whether the mobile is in active or idle state, consumes a lot of battery and loads the network with signaling messages. This constitutes a problem since mobiles have a limited capacity battery and since the wireless bandwidth is a real bottleneck.

  The paging is an efficient solution to this problem. With the paging solution, the network is divided into paging areas. The network does not have to keep exact location information of idle mobile nodes, but only information about the approximate location. Idle mobile nodes only have to notify the network of a change of paging area and not upon each handover. Thus, paging is an important issue for the mobility management to be considered when studying a micro-mobility protocol.

- Robustness. Future wireless networks are expected to support millions of customers. Robustness will be a major concern for such wireless networks where micro-mobility proposals must be able to handle the expected load with appropriate mechanisms.

- Intra-network traffic. One important issue is the traffic exchanged between two mobiles moving within the same network. This intra-network traffic must pass through the optimal path, in order to prevent the bandwidth waste which constitutes a problem in case of high traffic. Routing optimization can improve the QoS performance of the classes of service that generate this intra-network traffic.

- Load balancing. An efficient micro-mobility protocol performs load balancing: the protocol dynamically distributes the load in order to prevent nodes overload and network congestion.

3.6.1 Interaction with Mobile IP

In Cellular IPv6, the gateway router is not a foreign agent as in the case of Cellular IP defined for Mobile IPv4 [CGW$^+$00]. The mobile node in Cellular IPv6 networks is identified by its care-of address. This MN invokes the IPv6 Stateless address autoconfiguration mechanism to generate this CoA. Thus, the MN processes all IPv6 packets that it receives.

In HAWAII, each mobile node is assigned a co-located care-of address from the address space of the visited HAWAII network. Thus the Mobile IPv4 tunnel ends at the mobile node itself.

Concerning HMIP, it interacts with Mobile IPv4. A packet arriving at the top level of the hierarchy will be decapsulated and re-encapsulated before being re-routed to the next lower Regional FA in the hierarchy. This re-routing occurs at each level of the hierarchy, until the packet reaches the last point which is either the mobile node itself (in case of a co-located address) or a foreign agent. The latter can deliver the packet to the mobile node with no further special Mobile IP handling.

HMIPv6 supports two modes, the Basic mode and the Extended mode. In Basic mode, the RCoA is autoconfigured by the MN. Thus, packets are intercepted by the MAP and encapsulated again to the MN. On the other hand, if Extended mode is used, the RCoA is assigned to one of the MAP's interfaces. Thus, the MAP decapsulates the packets and then encapsulates them again to the MN. Hence, only one additional IPv6 header is needed for packets in Extended mode as opposed to two headers in the Basic mode.

With Fast Handovers and under normal conditions (i.e. not in a handover situation), the MN receives packets routed to its current CoA as specified in the Mobile IPv6 specification.

When a mobile moves to a nAR, the current CoA becomes an old CoA. The oAR intercepts any IPv6 packets addressed to the old CoA on the MN's old link, and routes each intercepted packet to the MN. If the MN is able to obtain a new CoA, the packets are routed to the nCoA of the MN on the new link. Otherwise, the packets are tunneled to the nAR address and the nAR decapsulates and delivers the packets to the MN via a Link Layer mechanism.

3.6.2 Protocol Layers

Cellular IP defines nodes working with advanced Layer 2 switch capabilities. These nodes must support delay devices and mobility management.

HAWAII uses classical IP routers that maintain routing tables and at the same time support the mobility management.

HMIP uses classical Mobile IP FA that must be capable of encapsulating IP packets: they operate at "Layer 3.5". HMIPv6 uses classical routers. The MAPs in HMIPv6 network must operate at Layer 3.5 because they must encapsulate IP packets. The MAPs can be at any level in the hierarchical networks of routers. Thus the routers in the HMIPv6 network are classical routers operating at Layer 3.5.

Fast Handovers protocol for Mobile IP v6 uses classical routers that encapsulate IP packets and supports MAP operating at Layer 3.5 when Fast Handovers protocol is integrated in HMIPv6 [SCEMB01].

3.6.3 Intra-network Traffic

In this section, we focus on the traffic between the MNs connected to the same network.

With Cellular IP [CGW$^+$00], the traffic coming from a MN must pass through the gateway even if the MN is communicating with another mobile in the same wireless network. This kind of routing increases the processing load on the gateway and on the neighboring stations. Route optimization is proposed in [SMG$^+$01] in order to route the intra-network traffic through the optimal path.

HAWAII works over IP. Thus the intra-network traffic will benefit from the classical IP routing.

With the Hierarchical Mobile IP, the intra-network traffic is directed to the HA of the destination. Thus, it must pass through the gateway before arriving to the corresponding node.

HMIPv6 benefits from the route optimization that exists in Mobile IPv6. Thus the intra-network traffic follows the optimal path. An important advantage of HMIPv6 is that MAPs can perform load balancing. In fact, each MAP is able to change its preference value. This value indicates the preference of the MAP to be used. A preference value equal to 15 means that the MAP cannot be used. The MAP changes its preference value dynamically based on the local policies and node overload.

As for Fast Handovers protocol, when the mobile node acquires a new care-of address in its new subnet, the intra-network follows the optimal path. In fact, the Fast Handovers protocol inherits the route optimization that exists in Mobile IPv6. However, during handovers, the problem is quite different. In fact, with the BETH mechanism, the mobile is moved to its new access router without the involvement of the Layer 3. As a result, MN continues to receive its packets via the established tunnel without changing its CoA. MN may delay obtaining a new CoA until its traffic is reduced or until the MN moves slowly enough that it has the time to obtain a new CoA. This approach makes the Route optimization difficult to achieve.

3.6.4 Paging

Hierarchical Mobile IP with its paging extension [HM00], Cellular IP and HAWAII support the passive connectivity and paging feature. They use the classical cellular telephony concepts of paging areas. As in GSM networks, the mobiles are grouped in paging areas and a router must perform paging to find the actual location of the MN in the network.

Paging extensions for Hierarchical Mobile IP are presented in [HM00] allowing idle mobile nodes to operate in a power saving mode while located within paging areas. Paging areas are sub-trees of the same hierarchy. In each paging area, the root of the sub-tree is called Paging Foreign Agent (PFA) and maintains a specific visitor's list with an idle flag set for each idle mobile in the area. After receiving a packet addressed to a mobile node located in a foreign network, the home agent tunnels this packet to the PFA. The PFA then pages the mobile node to re-establish a path towards the current point of attachment. Paging a mobile node can take place using a specific communication time-slot in the paging area similar to the paging channel in second generation cellular systems. In this case, the mobile node only needs to wake-up at predefined time intervals to check for incoming paging requests. Paging schemes increase the amount of time a mobile node can remain in a power saving mode.

Paging in Cellular IP does not require that all network nodes maintain Paging caches. Nodes that do not maintain a Paging cache simply broadcast packets destined to a mobile node that is not listed in their Route cache. Thus paging areas in Cellular IP depend on the network topology. In Cellular IP, paging is done with data packets themselves.

HAWAII uses IP multicast for paging. When paging is initiated at a node, the packet to be sent is buffered in the node and a paging message is transmitted to awake the mobile. The buffered packet is then delivered after that the mobile node activates a route leading to it,

HAWAII defines an algorithm to dynamically balance the paging load among the nodes of the network. Based on the current load of each node, a particular node is chosen to perform the paging each time it is needed.

3.6.5 Robustness

HMIP, HMIPv6, Cellular IP and HAWAII use a tree-like architecture having a gateway as its root. The gateway is the most loaded node in the network, processing all control messages and maintaining table entries for all MNs within the network. A direct consequence is that these architectures that rely on specific nodes like the gateway are weak.

In the case of Cellular IP and HMIP with paging, the paging mechanism increases the weakness since only few nodes maintain the paging information. This makes the network extremely vulnerable to the crash of these nodes.

On the other hand, HAWAII distributes the paging information inside the network and dynamically assigns the paging processing. This may help HAWAII to be more robust than other micro-mobility protocols.

In order to manage link failures, HAWAII and Cellular IP employ soft-state routing and paging entries. It is noteworthy that HMIPv6, HMIP and HAWAII work on the top of IP and thus benefit from the existing IP recovery mechanisms, in case of a station crash.

Tables 3.3 and 3.4 present the summary of the comparative study done in the previous subsections.

Protocol	**Cellular IP**	**HAWAII**
OSI Layer	L3 at all CIP nodes	L3 at all routers
Paging	implicit	explicit
Load Balancing	no	yes for paging
Robustness	Vulnerable to the crash of GW Vulnerable to the crash of nodes with Paging caches	Vulnerable to the DRR crash Assigns dynamically the paging information
Intra-Network Traffic	Non Optimal Path	Benefits from classical IP routing

Table 3.3: IP Micro-mobility protocols Comparison

Protocol	**HMIP**	**HMIPv6**	**Fast Handovers**
OSI Layer	L3.5 at all FAs	L3.5 at routers	L3.5 at ARs
Paging	explicit	no	no
Load Balancing	no	yes	no
Robustness	Vulnerable to GFA crash Vulnerable to PFA crash	Vulnerable to MAP crash	
Intra-Network Traffic	Non Optimal Path	Optimal Path	Optimal Path Non Optimal Path with BETH

Table 3.4: IP Micro-mobility protocols Comparison(bis)

3.7 Conclusion

In this chapter, we presented different IP mobility protocols that have been designed and implemented over the past several years. Micro-mobility protocols complement Mobile IP with fast, seamless and local handover control.

We exposed a comparative study of the micro-mobility protocols, taking into consideration major handover management issues. We can easily see, at the end of this comparison that each proposal has both strengths and weaknesses. It is difficult to claim that one protocol is better than another one. The choice of a micro-mobility protocol depends on the priorities imposed by the network management.

A number of open issues still remain. Micro-mobility protocols will have to support multiple classes of service including best effort and real-time classes. Thus, efficient micro-mobility protocols must provide QoS to the mobile nodes. Architectures for supporting QoS guarantees on the Internet can be broadly classified into two groups: the Integrated Services (IntServ) and the Differentiated Services (DiffServ). Both approaches suffer from drawbacks when used to provide resources for mobile nodes in wireless cellular environments. In fact, these approaches are designed in a static environment and as a result, they are not fully adapted to mobile environments, especially when Mobile IP is used as the mobility management protocol. In chapter 7, we present a survey of the works dealing with the QoS in micro-mobility domains. Moreover, we present our QoS model in Cellular IPv6 networks.

In the previous chapters, we have investigated the different mobility issues presented in the literature. At this stage, we present in the following chapters our propositions concerning the mobility handling and the QoS provisioning in wireless networks. In the next chapter, we start with a design of a multiservice call admission control supporting four classes of service in wireless multiservice mobile networks.

Chapter 4

Multiservice CAC and Handover Handling in Wireless Mobile Networks

4.1 Introduction

Call Admission Control (CAC) is a key element for ensuring the Quality of Service in mobile wireless networks. With the advances in wireless communication technology and the growing interest in deploying multimedia services in the wireless mobile networks, the issue of providing an efficient CAC has come to the fore. A suitable CAC for the cellular multimedia service networks is expected to make efficient use of the scarce wireless resource while supporting different services with different QoS metrics.

In this chapter, we propose a CAC model with three priority levels supporting four classes of service in a wireless mobile network. The idea is to design an efficient call admission control that deals with voice and data calls. These calls can be new (or originating) calls as well as handover calls. The proposed model manages to establish priority between voice and data calls, and between new and handover calls. It is assumed that the system operates under a reservation channel scheme and a queuing strategy in order to maintain the handover (HO) priority. Two different queuing disciplines have been introduced to further enhance the scheme characteristics. Our work proceeds in three consecutive phases.

In the first phase, the analytical approach has been adopted in order to evaluate the model performance where a three-dimensional Markov chain with discrete states and continuous time was elaborated in order to model the system. The mathematical tractability of the model being a very complex task, we shall present in the present chapter the analytical modeling procedure, pointing out the simplifying assumptions made to reduce the complexity.

The second phase consists of the validation of the analytical model. Hence, simulation was carried out in order to study the model performance and to compare simulation results with analytical ones.

In the third phase, we generalized our assumptions by taking a Web session model for data traffic. Two types of channel allocation were considered: circuit-switching for voice and burst allocation for data.

The chapter is organized as follows. In the next section, a summary of the recently published adaptive QoS provisioning techniques will be presented. This helps afterwards in describing our proposed model. Next, the multiservice CAC is introduced in section 5.4.5. Section 4.4 describes the parameters and assumptions adopted in order to reduce the complexity of the study. The detailed description of the analytical approach appears in sections 4.5 and 4.6 and its performance results in section 4.7. The model validation together with the assumptions generalization are presented in section 4.8 before concluding the chapter.

4.2 CAC in the literature

A suitable CAC for wireless mobile networks is expected to make efficient use of the scarce wireless resource while supporting different services with different QoS metrics. These QoS requirements dictate that the users must be accorded different priorities. In the literature, there has been intensive research on the design of CAC policies.

In a multiservice system, an important issue is the capacity allocation to each class of service. Two access policies might be the complete sharing (CS) and the complete partitioning (CP). The CS policy allows all users equal access to the bandwidth available all the time. This results in maximum usage of the available bandwidth. At the same time, it does not differentiate between users of different priorities which is problematic from a QoS perspective. The CP policy divides up the available bandwidth into separate sub-pools according to user types. This policy allows more control of the relative blocking/dropping probabilities at the expense of overall usage of the network capacity.

In [ES95], the authors introduce the hybrid policies. These policies seek to provide a compromise between the different access policies by subdividing the available bandwidth into sections. Part of the bandwidth is completely shared and the other part is completely partitioned. This allows more flexibility in catering the QoS requirements of the different user types while maintaining higher network usage.

The Virtual Partitioning with Priority (VPP), proposed in [CL00], combines the advantages of CS and CP where it provides high resource utilization with light offered load and fairness with heavy offered load. VPP is a state dependent trunk reservation scheme: if the current capacity usage by a specific class is greater than the nominal allocation then it is declared to be overloaded and a trunk reservation mechanism gives it lower priority in the admission of new calls.

In [KKC00], the authors consider multiple classes of adaptive multimedia services and the prioritization among classes in the bandwidth adaptation was assumed.

An efficient CAC must maintain a balance between two conflicting requirements: maximize the resource utilization and minimize the forced handover call dropping rate. In order to maintain the maximum resource utilization, the maximum number of calls should be admitted into a network, which may result in unacceptably high handover call dropping rates due to insufficient resources for handover calls. Therefore, it is very important to develop a *dynamic* CAC which estimates future resource demands, reserves the minimum amount of necessary resources to maintain an acceptable handover call dropping rate and provide high resource utilization.

In order to maximize the utilization efficiency of the wireless network and minimize the handover call dropping rate, a dynamic distributed CAC is implemented in [NS96] and [WWL98].

In fact, the base station makes an admission decision by exchanging information periodically with adjacent cells.

The admission threshold for different classes of traffic is calculated in [NA95] based on the traffic, the HO characteristics, the call holding time statistics and the desired QoS of each class of traffic. The admission threshold of different classes is recalculated periodically every fixed admission period. During this admission period, admission thresholds of different classes remain fixed and each BS estimates the Erlang load of each class of traffic. These estimates are reported to the network call controller at the end of the fixed admission period and are used by the network admission controller to calculate admission thresholds for different classes of traffic in the next fixed admission period.

In [McM95], the author proposes two buffers for the new calls and the handover calls. The model includes time-out so that if the waiting time in each queue exceeds a threshold, the call will leave the system. In this work, the handover and the new calls are assumed to be independent of each other.

In [CS00], the authors extend the previous model by taking into account the dependency of handover rates and the new calls. They also analyze the hysteresis controls according to the buffer occupancy of the new calls. In these schemes, the authors do not distinguish between data and voice traffic.

In order to study the performance issues of an integrated voice/data scheme, some works have implemented analytic models ([LZMF98], [ZA00]). The priority of voice handover requests over data handover requests is either given by reservation of channels exclusively for handover calls [ZA00], or is preemptive [LZMF98]. In [ZA00], there is no distinction between the new voice calls and the new data calls. In our study, we have prioritized new voice calls over new data calls.

Many papers considered an exponential data length. Exponential assumptions for arrival processes have well served telecommunication traffic theory for many years but it is interesting nevertheless to extend this assumption for data traffic: the latter is quite questionable and needs further generalization. Hence, we have adopted a real model for data traffic in the simulation in the third phase of our work.

4.3 CAC Proposed Model

Figure 4.1 illustrates the proposed CAC model. We consider a single cell in an homogeneous system comprising multiple cells. Each one is assumed to have an unchanging bandwidth (or channels) which provides service to heterogeneous users (data and voice users) who request service either as New or HO users. We propose the prioritization among four classes of service. Calls are first distinguished between New Calls (NC) and HO Calls. They are further divided into voice and data calls.

Since the breakdown of a call in progress called forced termination is less desirable than the blocking of a new call, a HO call has the priority over a NC. Since voice calls are delay sensitive, it is also assumed that voice calls have the priority over data calls. Thus, our model proposes a three priority-level scheme for the four classes in the following ascending order: New data call (NC-Data) / New voice call (NC-Voice) / Data handover call (HO-Data) / Voice handover call (HO-Voice).

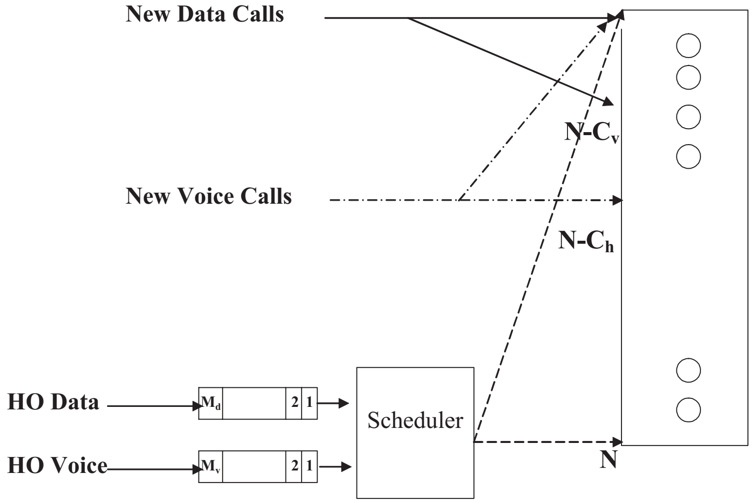


Figure 4.1: CAC Model

Before going into details of the proposed model, let us suppose that:

- C_h is the number of guard channels used exclusively by the HO requests.
- C_v is the number of channels that should not be used by new data calls such that $C_v \geq C_h$.
- N is the number of available channels per cell.
- M_v and M_d are respectively the voice and data queues capacity.

Our CAC model introduces two thresholds $N - C_v$ and $N - C_h$. A NC-data (respectively a NC-voice) is served only if the number of available channels is larger than C_v (respectively C_h). A HO request is served when finding an available channel (Fig. 4.1). Otherwise, the request is queued.

In each cell, there are two queues for voice and data handover requests. Since the voice and data handover requests can have access to all channels, we propose to apply two different scheduling policies in order to prioritize the voice HO requests.

- Due to its simplicity, many networks use the First In First Out (FIFO) policy. To allow some differentiation, some networks use the Head of the Line (HOL) with non-preemptive priority control scheme which schedules real-time HO traffic first. However, this scheme degrades severely the performance of low priority traffic as the portion of high priority traffic increases.
- The Queue Length Threshold (QLT) gives priority to non-real time traffic whenever the queued non-real time HO requests are above some threshold. It gives a better overall tradeoff between performance and complexity.

A queued voice handover request is deleted from the queue when it passes through the handover area before getting a channel or if its communication is completed before passing through the handover area.

Data connections typically do not require real-time transport of the information through the network and are therefore more tolerant to delay as compared to voice connections. Thus, we suppose that there is no handover area for data users, but there is a cell boundary between two neighboring cells. As a result, whenever a data handover request is not satisfied within the current cell, it will be transferred to the data queue of the target cell instead of dropping it. In fact, deleting the handover request might impose a high penalty incurred to re-initiate the call and retransmit a large stream of data. Consequently, the average delay that is computed in our study is the average waiting delay of the data HO requests. Even though data calls are delay tolerant, they can encounter high delays: in fact they can be transmitted from a queue in an old cell to another queue in a new cell.

At this point, we will focus on the analytical model of the proposed CAC. This analytical model will help us to evaluate the performance of the CAC. Afterwards, we will compare the analytical results with those obtained from simulations. Note that in the analytical model, we consider that data and voice calls need one channel in order to simplify the analysis. In the simulation presented in section 4.8, data calls require more resources than the voice calls.

4.4 Assumptions and Parameters

Before presenting the analytical model, we make in this section some assumptions needed for the model.

The unencumbered session duration of a voice call T_{cv} is assumed to be exponentially distributed with mean $1/\mu_{cv}$. The data length T_{cd} is also supposed to be exponentially distributed with mean $1/\mu_{cd}$. Let T_{dwell} be the mobile users dwell time in a cell, assumed to be exponentially distributed with mean $1/\mu_{dwell}$.

An occupied channel will be freed by an user either at the end of the call (i.e. after the expiration of the session duration) or after a handover (i.e. after the sojourn in the current cell). As a result, the channel holding time of a voice (or data) call is equal to the smaller value between T_{dwell} and T_{cv} (or T_{cd}). Therefore, the channel holding time of voice and data calls is exponentially distributed with means $E[T_v] = 1/\mu_v$ and $E[T_d] = 1/\mu_d$ respectively such that:

$$E[T_v] = \frac{1}{\mu_v} = \frac{1}{\mu_{cv} + \mu_{dwell}} \tag{4.1}$$

$$E[T_d] = \frac{1}{\mu_d} = \frac{1}{\mu_{cd} + \mu_{dwell}} \tag{4.2}$$

The dwell time of the mobiles in the handover area T_w is assumed to be an exponentially distributed random variable with mean $1/\mu_w$.

We suppose that the arrival processes of new voice calls and new data calls are Poisson with rates λ_{nv} and λ_{nd} respectively.

In order to analyze the system, the handover arrival rates must be computed. The arrival processes of voice HO requests and data HO requests are assumed to be Poisson with mean rates λ_{hv} and λ_{hdata} respectively.

The Poisson assumption for HO traffic was made in order to make the analysis tractable. In [CL95], the authors showed that in a *blocking* wireless network, the Poisson model for handover traffic is reasonable although not exact. In fact, it gives good results in comparison with the simulation results over the entire speed range. Simulation results presented in section 4.8 reflect concordance between the analytical and the simulated model. As a conclusion, we can claim that the simplifying hypothesis adopted has a slight affect on the results.

Assuming an equilibrium homogeneous mobility pattern, the mean number of incoming users into a cell is equal to that of outgoing ones from the cell. Hence, the arrival rate of HO requests is equal to the departure rate from the cell. Consequently, similar to [ZA00], we have the following relation:

$$\lambda_{hv} = E[C_v]\mu_{dwell} \tag{4.3}$$

where $E[C_v]$ is the average number of voice users holding channels in a cell.

For data users, the rate of calls going out of a cell without completing communication is $E[C_d]\mu_{dwell}$, where $E[C_d]$ is the average number of data users holding channels in a cell. If the system has homogeneous cells, the arrival rate of data handover request calls is given by:

$$\lambda_{hdata} = E[C_d]\mu_{dwell} \tag{4.4}$$

A data handover request is transferred from the queue of the current cell to the queue of the target cell when it moves out of the cell before getting a channel. If L_d is the data queue average length, the transfer rate of data HO requests from one queue to another is:

$$\lambda_t = L_d\mu_{dwell} \tag{4.5}$$

Thus, we define a new random variable λ_{hd} by:

$$\lambda_{hd} = \lambda_{hdata} + \lambda_t = (E[C_d] + L_d)\mu_{dwell} = N_d\mu_{dwell} \tag{4.6}$$

where N_d is the average number of data handover requests in a cell.

4.5 System Analysis

The state of a cell is defined at time $t \in IR^+$ by the vector (X_t, Y_t, Z_t), where X_t, Y_t and Z_t are three random variables such that:

- X_t is the sum of channels used by voice calls and the number of voice handover requests in the voice handover queue.
- Y_t is the number of channels used by data calls.
- Z_t is the number of data handover requests in the data handover queue.

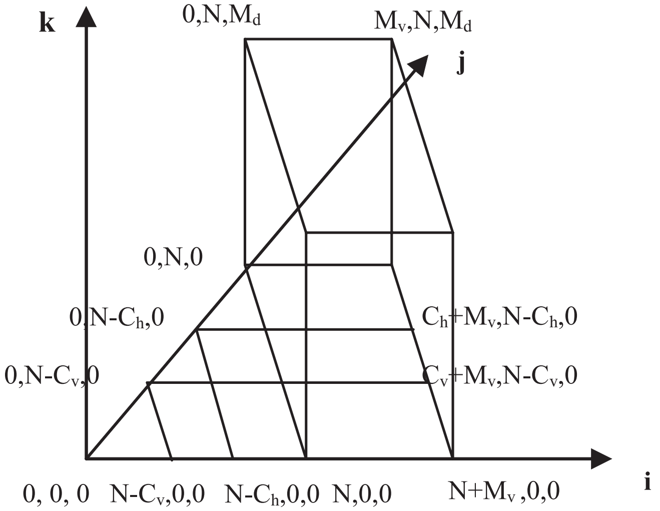


Figure 4.2: State Space

The arrival process of new voice calls and new data calls is a Poisson process. On the other hand, the channel holding time of voice and data calls has an exponential distribution. Moreover, a request in a HO queue will leave the queue either after the expiration of the dwell time in the handover area (exponentially distributed), or at the end of the communication of the waiting HO request (exponentially distributed) , or at the end of the service of an ongoing call (exponentially distributed). Thus, the waiting time of a handover request is exponentially distributed.

The conditional expectation $E(X_{t+\delta t}, Y_{t+\delta t}, Z_{t+\delta t} | X_t, Y_t, Z_t)$ is independent from the sigma algebra $\sigma(X_u, Y_u, Z_u; u < t)$ where $\delta t \in I\!R^+$, due to the memoryless property of the exponential distribution.

As a result, the vector (X_t, Y_t, Z_t) is a Markov chain (birth and death process) with discrete states and continuous time. Furthermore, this Markov chain is irreducible (see the graphs associated with the infinitesimal generator of the chain depicted below) and has a finite state space (see figure 4.2). Therefore, this chain is ergodic. Thus, when t tends towards ∞, the Markov chain (X_t, Y_t, Z_t) tends towards a stationary state (X, Y, Z).

We suppose that X, Y and Z take respectively the values i, j and k such that:

- $i \in [0, .., N + M_v]$
- $j \in [0, .., N]$
- $k \in [0, .., M_d]$

We are interested in the computation of the stationary probabilities. Thus, we have to resolve the Kolmogorov equations in the stationary state. These equations are given by $\mu A = 0$, where A is the infinitesimal generator of the Markov chain and μ is the stationary distribution [DKKT00].

The Kolmogorov equations being difficult to write down, we shall in a first step illustrate the state space of the random variables. In a second step, we will graphically represent these equations that we solve iteratively.

In figure 4.2, we depict the state space. This diagram comprises Ns states. In order to have an idea about the system complexity, we have computed Ns: in fact, $Ns = (N+1)[N+2+2.(Mv.Md+Mv+Md)]/2$.

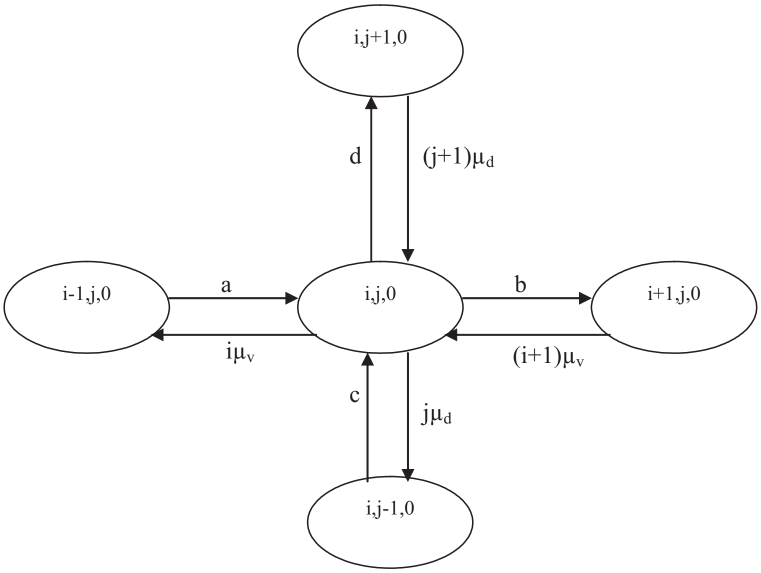


Figure 4.3: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j < N$

We have drawn the Markov chain related to the two types of scheduling. For the HOL scheme, we distinguish three cases:

1. $i+j < N$: this case is illustrated by figure 4.3. As we can see, k equals zero because the total number of busy channels does not exceed the cell capacity. Consequently, the data HO requests can be satisfied by the available channels. In order to simplify the representation, we define λ_v and λ_d by $\lambda_v = \lambda_{nv} + \lambda_{hv}$ and $\lambda_d = \lambda_{nd} + \lambda_{hd}$.

   As new data calls (respectively new voice calls) are blocked when the number of busy channels exceeds $N - C_v$ (respectively $N - C_h$), it is not hard to compute the transition rates a, b, c and d as follows:

$$a = \begin{cases} \lambda_v & if\; 0 \leq (i-1+j) < N - C_h \\ \lambda_{hv} & if\; (i-1+j) \geq N - C_h \end{cases}$$

$$b = \begin{cases} \lambda_v & if\; 0 \leq (i+j) < N - C_h \\ \lambda_{hv} & if\; (i+j) \geq N - C_h \end{cases}$$

$$c = \begin{cases} \lambda_d & if\; 0 \leq (i+j-1) < N - C_v \\ \lambda_{hd} & if\; (i+j-1) \geq N - C_v \end{cases}$$

$$d = \begin{cases} \lambda_d & if\; 0 \leq (i+j) < N - C_v \\ \lambda_{hd} & if\; (i+j) \geq N - C_v \end{cases}$$

2. $i+j=N$: this case is depicted in figure 4.4. Note that in this figure, $a = 1 \ for \ k = 0$ and 0 otherwise

3. $i+j>N$: the states of this case are exhibited in figure 4.5.

As for the QLT scheme, we distinguish two cases as well:

1. $i+j<N$: the policies HOL and QLT are applied whenever the number of busy channels exceeds N. Consequently, the states of this case are the same in both HOL and QLT schemes (figure 4.3).

2. $i+j=N$: this case can be divided into two sub-cases depending on whether $k < L_{th}$ or $k \geq L_{th}$.

   The first one ($k < L_{th}$) is similar to the HOL scheme with $i+j=N$. In fact, whenever the data length is less than the threshold L_{th}, priority is affected to the voice HO requests. Consequently, in this case, QLT behaves similarly to the HOL scheme (figure 4.4).

   As for the second sub case ($k \geq L_{th}$), the states are illustrated in figure 4.6.

3. $i+j>N$: this case can be split up into two sub-cases as well depending on whether $k < L_{th}$ (figure 4.7) or $k \geq L_{th}$ (figure 4.8). Note that in figure 4.7, the transition rates A and B are such that:

$$A = \begin{cases} (k+1)\mu_{dwell} & if \ (k+1) \leq (L_{th}-1) \\ j\mu_d + (k+1)\mu_{dwell} & if \ ((k=0, L_{th}=1) \ or \ (k = L_{th}-1)) \end{cases}$$

$$B = \begin{cases} (i+1)\mu_v & if \ ((k=0, L_{th}=1) \ or \ (k = L_{th}-1)) \\ 0 & otherwise \end{cases}$$

The space of the states is finite. On the other hand, the graphs associated with the infinitesimal generator show that the Markov chain is irreducible. Thus as stated before, the Markov chain is ergodic and has a stationary probability distribution [DKKT00].

For each scheduling policy and given the normalizing condition, we obtain Ns independent equations used to derive all the stationary probabilities $P(i,j,k) \ for \ i = 0,..,N+M_v; j = 0,..,N; k = 0,..,M_d$.

The problem that we had to solve is that in these independent equations, the HO rates are unknown input values. To be more precise, these HO rates are expressed as a function of the parameters $E[C_v]$, $E[C_d]$ and L_d as depicted in equations 4.3, 4.4, 4.5 and 4.6. The above-mentioned parameters are in their turn function of $P(i,j,k)$ such that:

$$E[C_v] = \sum_{i=1}^{N-1} i \sum_{j=0}^{N-i-1} P(i,j,0) + \sum_{j=0}^{N} \sum_{i=N-j}^{N+M_v-j} (N-j) \sum_{k=0}^{M_d} P(i,j,k) \tag{4.7}$$

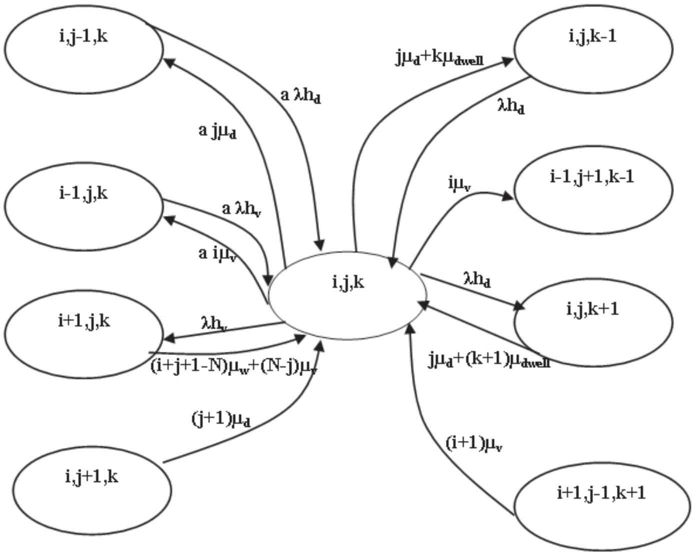


Figure 4.4: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j = N$ with HOL and QLT

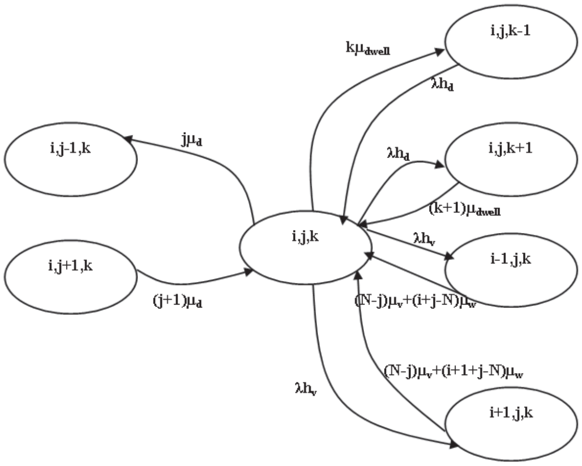


Figure 4.5: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j > N$ with HOL

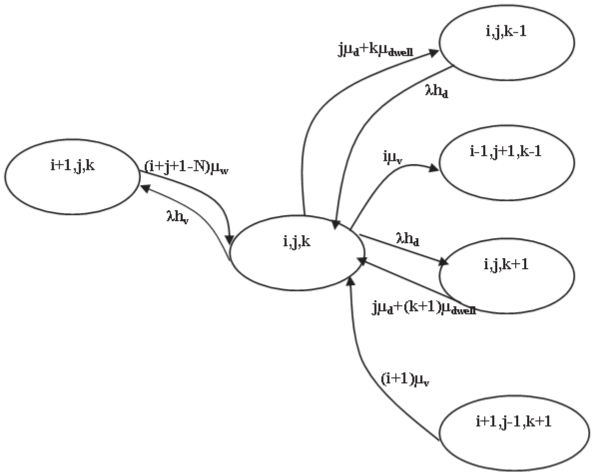


Figure 4.6: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j = N$ and $k \geq L_{th}$ with QLT

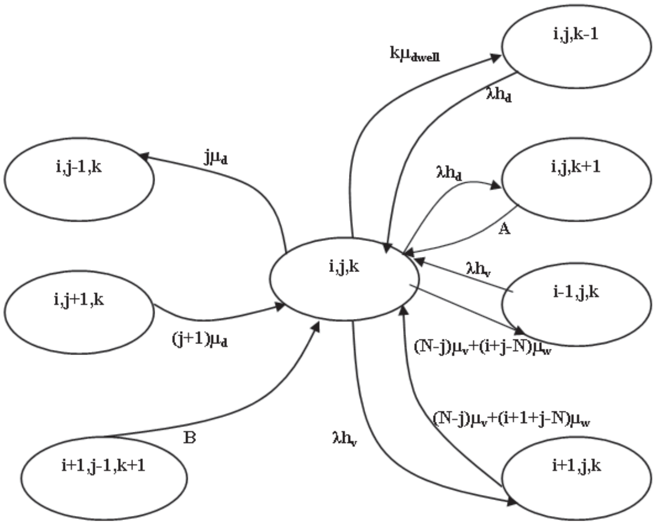


Figure 4.7: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j > N$ and $k < L_{th}$ with QLT

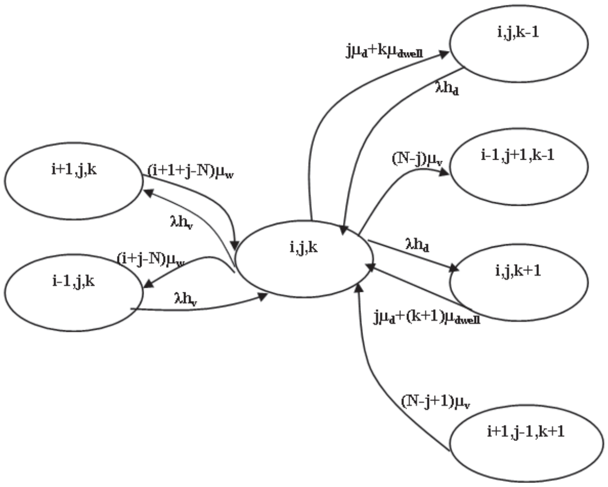


Figure 4.8: Graph Associated with the Infinitesimal Generator of the Markov Chain for $i+j > N$ and $k \geq L_{th}$ with QLT

$$E[C_d] = \sum_{j=1}^{N-1} j \sum_{i=0}^{N-j-1} P(i,j,0) + \sum_{j=1}^{N} j \sum_{i=N-j}^{N+M_v-j} \sum_{k=0}^{M_d} P(i,j,k) \tag{4.8}$$

$$L_d = \sum_{k=1}^{M_d} k \sum_{j=0}^{N} \sum_{i=N-j}^{N+M_v-j} P(i,j,k) \tag{4.9}$$

In order to compute the HO rates, we applied the iteration method which consists of the following steps.

1. First, we select arbitrary initial values for λ_{hv} and λ_{hd}.
2. Next, using the N_s independent equations, the stationary probabilities $P(i,j,k)$ are computed for the selected values.
3. Afterwards, we calculate $E[C_v]$, $E[C_d]$ and L_d using equations 4.7, 4.8 and 4.9.
4. At this point, we compute $new\lambda_{hv}$ by substituting $E[C_v]$ into equation 4.3. We also compute $new\lambda_{hd}$ by substituting $E[C_d]$ and L_d into equation 4.6.
5. If $\mid new\lambda_{hv} - old\lambda_{hv} \mid \leq \epsilon$ and $\mid new\lambda_{hd} - old\lambda_{hd} \mid \leq \epsilon$, where ϵ is a small positive number needed to check the convergence, then we stop executing the iteration method; the convergence is reached. Otherwise we re-compute $P(i,j,k)$ using the new HO rates and we re-execute the steps explained before, until convergence is reached.

This method helps to attain the HO rates convergence. When the convergence is reached, we derive the performance parameters as expressed in the following section.

We elaborated a program with Maple interfaced with Matlab in order to compute the stationary probabilities and the HO rates and to retrieve the performance parameters after the convergence of the iteration method. The accuracy of the results was checked by a simulation model as will be explained in section 4.8.

4.6 Performance Parameters

4.6.1 Voice Performance Parameters

For voice connections, the connection must be dropped if the mobile moves into a congested area where there is no wireless channel available. The voice QoS metrics are the forced termination probability, the blocking probability of NC-voice and the blocking probability of HO-voice request.

- A new voice call will be blocked if the total number of busy channels $(i+j)$ is greater than or equal to $N - C_h$. The blocking probability of a new voice call B_{nv} is:

$$B_{nv} = 1 - \sum_{i=0}^{N-C_h-1} \sum_{j=0}^{N-C_h-1-i} P(i,j,0) \tag{4.10}$$

- A voice handover request is blocked if the mobile leaves the handover area before being served. Calls that encounter a full queue receive continued support from the source gateway for as long as the call remains in the transition zone. Some of them may terminate before they are forced into termination. Hence, the blocking probability of a voice handover request is B_{hv} such that:

$$\begin{aligned} B_{hv} &= Prob(queuefull).Prob(T_{cv} \geq T_w) \\ &= \sum_{j=0}^{N} \sum_{k=0}^{M_d} P(N+M_v-j,j,k) \frac{\mu_w}{\mu_w+\mu_{cv}} \end{aligned}$$

  In our approach, we do not consider the eventuality of the call completion within the HO area in order to simplify the adopted assumptions. Thus,

$$B_{hv} = \sum_{j=0}^{N} \sum_{k=0}^{M_d} P(N+M_v-j,j,k) \tag{4.11}$$

- The time-out probability of a voice handover request is given by [ZA00] and is expressed as the ratio of the HO requests that do not get a channel within the HO area to the non-blocked HO requests.

$$P_{v-out} = \frac{\mu_w L_v}{(1-B_{hv})\lambda_{hv}} \tag{4.12}$$

where L_v is the average length of the voice queue such that:

$$L_v = \sum_{m=1}^{M_v} m \sum_{j=0}^{N} \sum_{k=0}^{M_d} P(N+m-j, j, k) \tag{4.13}$$

- The failure probability of a voice handover request B_{fv} is either due to a HO blocking or to a time-out event within the HO area. As a result, $B_{fv} = B_{hv} + (1 - B_{hv})P_{v-out}$

- The forced termination probability is defined as the probability that a call which was not blocked will be interrupted during its lifetime due to a HO failure. In order to derive this performance parameter, some probabilities must be computed first.

  Let P_{Hvoice} be the probability to require a handover.

$$\begin{aligned} P_{Hvoice} &= Prob(T_{cv} > T_{dwell}) \\ &= 1 - Prob(T_{cv} \leq T_{dwell}) \\ &= 1 - \int_0^{\infty} e^{-\mu_{dwell}\tau} \mu_{cv} e^{-\mu_{cv}\tau} d\tau \\ &= \frac{\mu_{dwell}}{\mu_{dwell} + \mu_{cv}} \end{aligned} \tag{4.14}$$

  Let a denote the probability that a call will make a handover attempt and will fail on that attempt. Thus,

$$a = P_{Hvoice} B_{fv}$$

  Let b denote the probability that a call will make a handover attempt and will succeed on that attempt. Thus,

$$b = P_{Hvoice}(1 - B_{fv})$$

  Consequently, the forced termination probability is as follows:

$$\begin{aligned} P_{ftvoice} &= \sum_{i=0}^{\infty} a.b^i \\ &= P_{Hvoice} B_{fv} \sum_{i=0}^{\infty} (P_{Hvoice}(1 - B_{fv}))^i \\ &= \frac{P_{Hvoice} B_{fv}}{1 - P_{Hvoice}(1 - B_{fv})} \end{aligned} \tag{4.15}$$

4.6.2 Data Performance Parameters

The data QoS metrics are the average waiting delay, the average queue length, the blocking probability of new data calls, and the blocking probability of HO requests.

- A new data call will be blocked if the total number of busy channels is greater than or equal to $N - C_v$. Consequently, the blocking probability of a data call B_{nd} is:

$$B_{nd} = 1 - \sum_{i=0}^{N-C_v-1} \sum_{j=0}^{N-C_v-1-i} P(i,j,0) \tag{4.16}$$

- A data handover request is blocked if the data queue is full. Therefore, the blocking probability of a data handover request B_{hd} is such that:

$$\begin{aligned} B_{hd} &= Prob(queuefull) \\ &= \sum_{j=0}^{N} \sum_{i=N-j}^{N+M_v-j} P(i,j,M_d) \end{aligned} \tag{4.17}$$

- The average waiting data delay is the delay encountered by the data HO requests in the data HO queue:

$$T_{delay} = N_h E[T_w] \tag{4.18}$$

  where :

  - $E[T_w]$, the average value of waiting time of a data handover request, is given by Little's formula [Kle75]:

$$E[T_w] = \frac{L_d}{(1 - B_{hd})\lambda_{hd}} \tag{4.19}$$

  - N_h is the average number of handovers encountered by a data handover request during its lifetime. Reference [ZA00] demonstrates that N_h is such that:

$$N_h = \frac{E[T_{cd}]}{\frac{N_d}{(1-B_{nd})\lambda_{nd}+(1-B_{hd})\lambda_{hd}} - E[T_w]} \tag{4.20}$$

4.7 Performance Results

For the analytical model, we consider one circular cell with radius R and we assume that users are pedestrians. The model parameters were chosen to study the system performance and the impact of some relevant parameters.

Parameters are set as follows: $R = 0.1Km$, user speed $E[V] = 1.8Km/h$, $E[T_{cv}] = 120s$, $E[T_{cd}] = 60s$, $E[T_{dwell}] = 314s$, $E[T_w] = 20s$, $N = 10, \epsilon = 10^{-8}$ and $\lambda_{nv} = \lambda_{nd} = \lambda_{ntot}/2$. Note that λ_{ntot} is the call originating rate density: it represents the number of new calls per second per kilometer square. The call originating rate density will be used in most of the simulations presented in the subsequent chapters.

4.7.1 Impact of HO queues

Figure 4.9 shows the performance parameters obtained with $(M_v, C_v, C_h) = (1, 1, 1)$ and with the call originating density equal to 2.5 $call/s/Km^2$. It can be seen that when the length of the HO data requests M_d grows, the probabilities B_{nv}, B_{nd}, B_{hv} and $P_{ftvoice}$ remain unchanged. On the other hand, the blocking probability of data HO requests B_{hd} decreases rapidly. As a matter of fact, the increase of M_d gives more chance to the data HO requests to be inserted into the data queue and thus to be satisfied. With the increase of M_d, there are more data HO requests in the data queue. Hence L_d and T_{delay} increase.

When $M_d \geq 2$, T_{delay} and L_d are almost constant. This is because the dwell time of handover calls is limited. Thus, even more chance can be affected to the data HO queue, the limited HO area prevents the requests to profit from the length increase.

It is to be noted that B_{hd} and B_{hv} cannot be compared because the handover rates obtained after convergence are quite different. There are respectively 0.013 $calls/sec/Km^2$ for voice users and 0.0065 $calls/s/Km^2$ for data users.

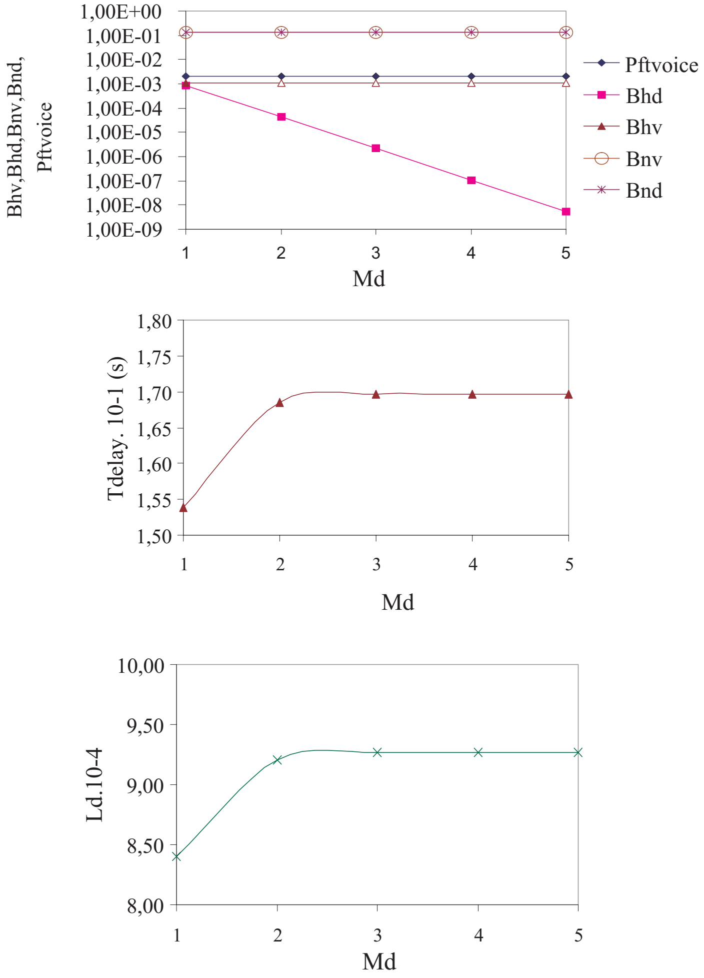


Figure 4.9: QoS Metrics versus M_d for $(M_v, C_v, C_h) = (1, 1, 1)$, λ_{ntot} with HOL

Some results concerning the voice queue length M_v have also been retrieved. We have obtained the same conclusion as for the impact of the data queue length.

B_{hv} decreases rapidly and $P_{ftvoice}$ decreases from $M_v = 1$ to $M_v = 2$. When $M_v \geq 2$, $P_{ftvoice}$ is almost constant. This is also related to the limited dwell time of handover calls in the HO area.

4.7.2 Impact of C_v and C_h

Figures 4.10 and 4.11 depict the voice and data QoS metrics with HOL for $(C_v, C_h) = \{(0,0),(1,0),(2,0)\}$.

When C_v increases, some priority is given to the new voice calls. Therefore, B_{nv} decreases and B_{nd} increases. Moreover, the forced termination probability of HO calls $P_{ftvoice}$, the average waiting delay T_{delay} and the average length of data queue L_d decrease. Blocking new data calls, when the number of busy channels reaches $N - C_v$, gives HO voice calls more chance to find free channels and reduces the number of HO requests waiting in the data queue.

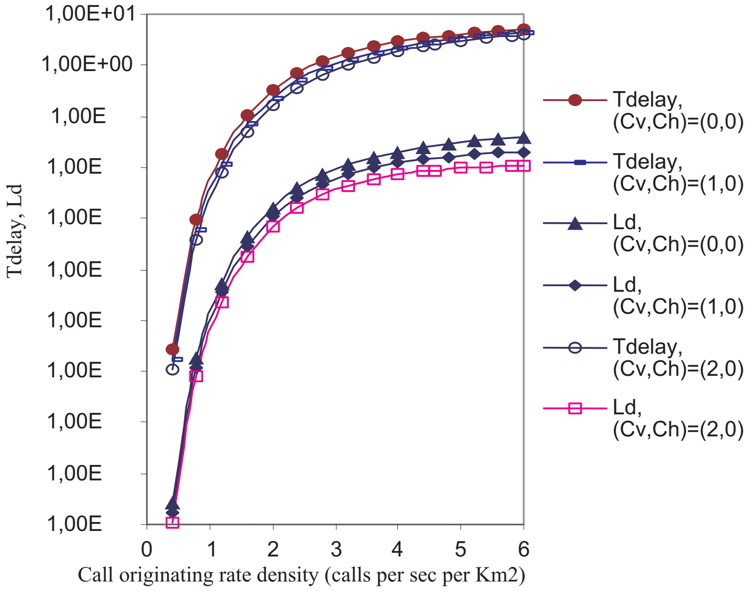


Figure 4.10: Impact of C_v: T_{delay} (s) and L_d versus call originating rate density for $(M_v, M_d) = (2,5)$ with HOL

When C_h increases (Figure 4.12), more priority is given to the HO calls and thus B_{nv} increases. This leads to the reduction of B_{nd}, $P_{ftvoice}$, T_{delay} and L_d. This reduction is accompanied with a small increase in B_{nv}. This exchange is important because forced termination is more critical than the new call blocking probability.

Table 4.1 shows the performance obtained using the QLT and HOL scheduling applied to the data and voice HO requests. There is a slight improvement of performance with $Lth = 2$, relative to the limited dwell time of HO calls.

In order to simplify the interpretation of the result, let us define: $\Delta X = \frac{X_{HOL} - X_{QLT}}{X_{HOL}}$ which denotes the relative variation of X a variable measured with HOL and QLT scheduling policies.

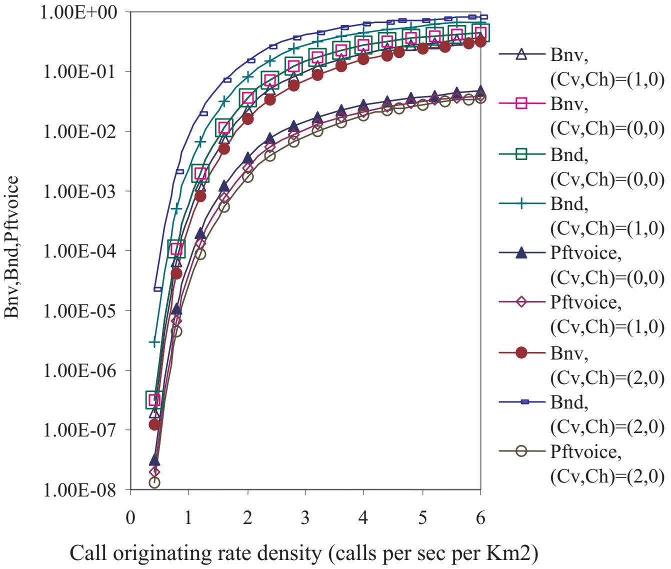


Figure 4.11: Impact of C_v: Call Blocking Probability versus call originating rate density for $(M_v, M_d) = (2, 5)$ with HOL

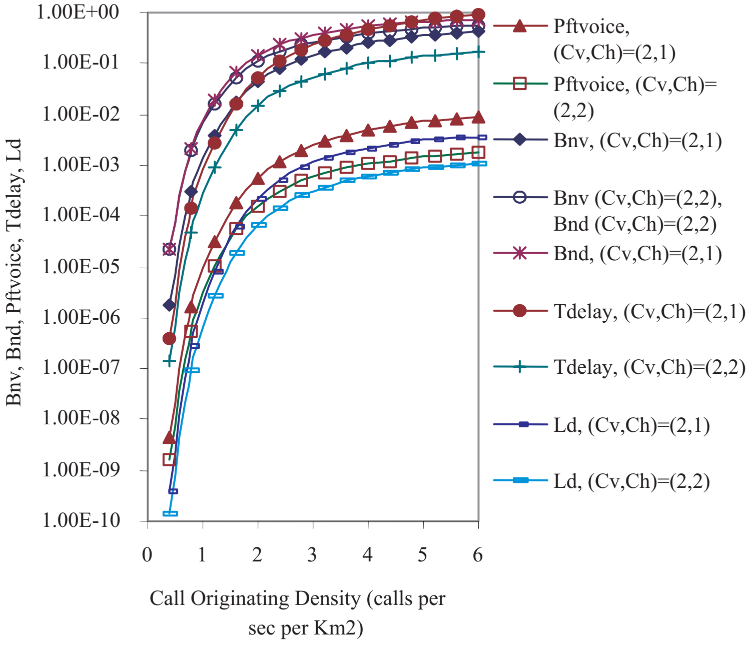


Figure 4.12: Impact of C_h for $(M_v, M_d) = (2, 5)$ with HOL

$\Delta X > 0$ means that QLT improves the performance of X, compared to HOL. This is valid because X represents either the average data waiting delay or the new call blocking or the forced termination probability or the average queue length. Thus the smaller the value of X is, the better QoS is.

The new call blocking probabilities measured with QLT do not change in comparison with HOL because QLT is applied on the HO requests only.

Since the priority is given to the data HO requests with the QLT when the data queue length reaches the threshold L_{th}, T_{delay} and L_d decrease whereas $P_{ftvoice}$ increases slightly. However, these results show that the reduction in T_{delay} and L_d is more significant than the $P_{ftvoice}$ increase with QLT. Table 4.1 shows that for $(C_v, C_h) = (2, 1), \Delta T_{delay} = \Delta L_d = 12.8\%$ and $\Delta P_{ftvoice} = -3.96\%$.

In fact, the data channel holding time is smaller than the voice channel holding time. Thus, even if the data HO request is served before the voice HO request, the channel will be busy for a time duration smaller than that of a voice HO call. Thus, the performance amelioration of data calls is made without inducing a perceptible degradation of the voice QoS.

At this point, let us compare the improvement of QLT with different values of (C_v, C_h).

Table 4.1 indicates that for $(C_v, C_h) = (2, 1), \Delta T_{delay} = \Delta L_d = 12.8\%$ and $\Delta P_{ftvoice} = -3.96\%$ whereas for $(C_v, C_h) = (2, 2), \Delta T_{delay} = \Delta L_d = 10.6\%$ and $\Delta P_{ftvoice} = -5\%$. As we can see, the QLT improvement is higher with $C_h = 1$ than with $C_h = 2$: the increase of $P_{ftvoice}$ (respectively the decrease of T_{delay} and L_d) with QLT is higher (resp. lower) when $C_h = 2$.

This phenomenon is explained by the fact that while C_h increases, we block more NC-voice calls and there are much less voice HO requests in the voice queue. We can obtain this result by comparing the ratio of the data average length to the voice average length L_d/L_v.

In fact, the numerical results indicate that L_d/L_v for $C_h = 2$ is greater than L_d/L_v for $C_h = 1$ (Table 4.1): the number of data HO requests exceeds the number of voice HO requests with the increase in C_h. Then, $P_{ftvoice}$ increases more with QLT when $C_h = 2$. Consequently, QLT performs better with $C_h = 1$. Similarly, we deduce from Table 4.1 that QLT performs better with $C_v = 2$.

4.8 Model Validation and Assumption Generalization

In order to examine the accuracy of the proposed analytical method, simulation results are compared with those obtained from the analytical approach. Computer simulations have been derived by assuming a seven cell network. The edges of the simulated space wrap around to the opposite edges with each cell having a complete set of interfering cells so as to avoid the border effect. Simulation was done using the discrete event simulation system OMNeT++ [OMN]. In

(C_v, C_h)	$(1, 1)$	$(2, 1)$	$(2, 2)$
$\Delta T_{delay}, \Delta L_d$	11.04%	12.8%	10.6%
$\Delta P_{ftvoice}$	-5.38%	-3.96%	-5%
$L_d/L_v, \lambda_{ntot} = 4calls/s/Km^2$	0.786	0.497	0.771

Table 4.1: Computation of $\Delta T_{delay}, \Delta L_d, \Delta P_{ftvoice} and L_d/L_v$

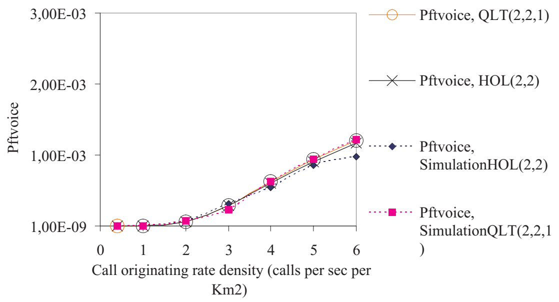


Figure 4.13: Comparison between simulation and analytical results: $P_{ftvoice}$ versus call originating rate density,$(M_v, M_d) = (2,5),(C_v, C_h) = (2,1),(2,2)$ with HOL

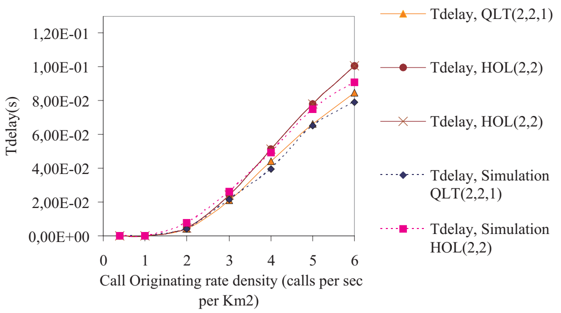


Figure 4.14: Comparison between simulation and analytical results: T_{delay}(s) versus call originating rate density,$(M_v, M_d) = (2,5),(C_v, C_h) = (2,1),(2,2)$ with HOL

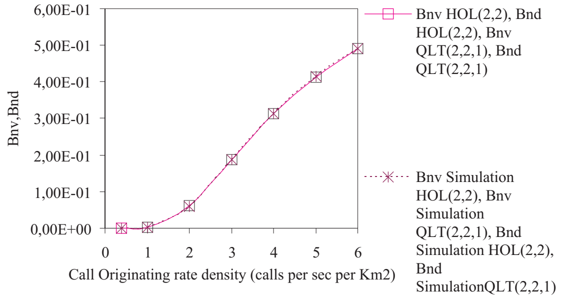


Figure 4.15: Comparison between simulation and analytical results: B_{nv}, B_{nd} versus call originating rate density, $(M_v, M_d) = (2,5),(C_v, C_h) = (2,1),(2,2)$ with HOL

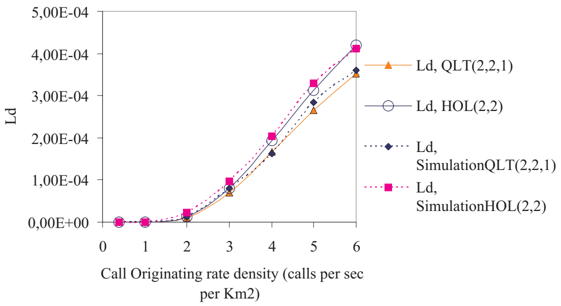


Figure 4.16: Comparison between simulation and analytical results: L_d versus call originating rate density, $(M_v, M_d) = (2,5),(C_v, C_h) = (2,1),(2,2)$ with HOL

a first approach, the data length is considered to be exponential in the simulation with mean 30s.

Figures 4.13, 4.14, 4.15 and 4.16 show a good agreement between simulation and analytical model over the range of traffic intensities investigated. This implies the accuracy of our analytical model. Note that the mean data length in the analytical results presented in the above-mentioned figures is assumed to be exponentially distributed with mean 30s.

In a second approach, we simulate our network with a real data model proposed by [Eur98].

The data traffic considered represents a typical WWW session (UDD 32kb/s) that consists of a sequence of packet calls. This data model was presented in chapter 1.

In order to benefit from the time where the HTTP user is inactive, we propose to apply in the simulation a statistical multiplexing for data traffic. Hence, the channel is allocated during the burst duration and is released during the reading time. In fact, the reading time might attain an important value (i.e. 412 s according to [Eur98]). Thus, even if the channel negotiation may require a non-negligible amount of time, we prefer to higher the bandwidth utilization by applying a statistical multiplexing than to block the unused channels for a great part of time. On the contrary, the channel is allocated to a voice user for the whole duration of the communication. Furthermore, we consider that a data (respectively voice) user needs two channels (respectively one channel) to be satisfied. Hence B_{nd} is much greater than B_{nv} (Figure 4.20).

Concerning HOL and QLT comparison, at high loads QLT improves the performance of the data traffic (by reducing T_{delay}, L_d) at the expense of a slight deterioration of the voice QoS (by slightly increasing $P_{ftvoice}$) (figures 4.17, 4.18, 4.19). The impact of QLT is better when C_h is smaller. This shows that QLT benefits are still valid when a generalized model is adopted.

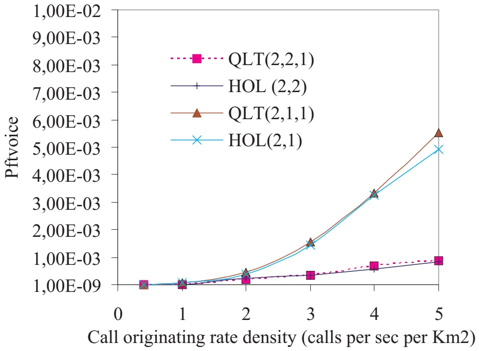


Figure 4.17: Model validation using Web model, $(M_v, M_d) = (2,5), (C_v, C_h) = (2,2)$: $P_{ftvoice}$ versus call originating rate density

4.9 Conclusion

QoS provisioning in wireless mobile networks supporting different classes of service requires efficient allocation of wireless resources. Our proposed CAC scheme is a three-priority level

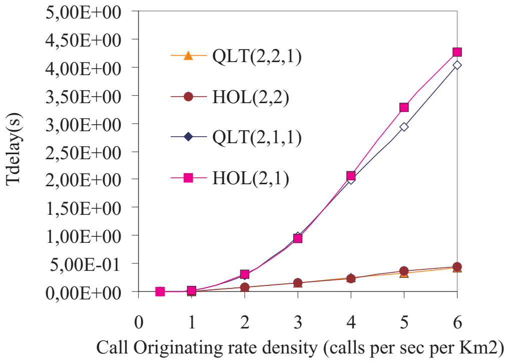


Figure 4.18: Model validation using Web model, $(M_v, M_d) = (2, 5), (C_v, C_h) = (2, 2)$: $T_{delay}(s)$ versus call originating rate density

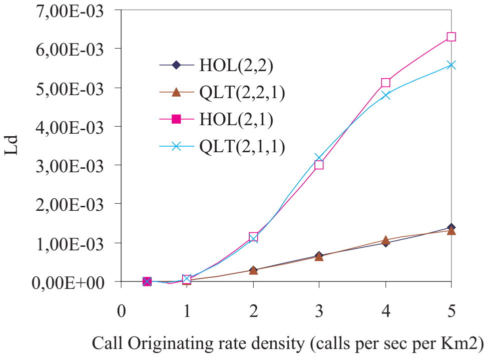


Figure 4.19: Model validation using Web model, $(M_v, M_d) = (2, 5), (C_v, C_h) = (2, 2)$: L_d versus call originating rate density

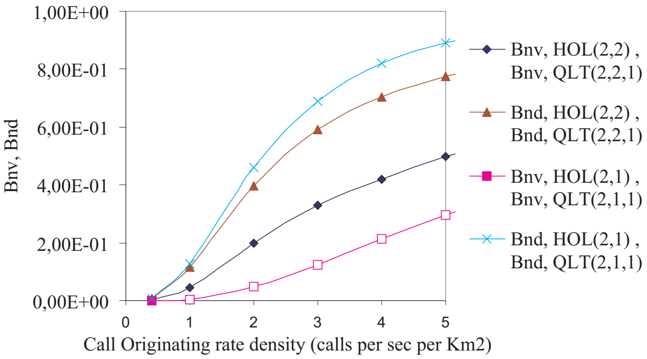


Figure 4.20: Model validation using Web model, $(M_v, M_d) = (2, 5), (C_v, C_h) = (2, 2)$: B_{nv}, B_{nd} versus call originating rate density

scheme that serves four classes of service. Two types of scheduling were envisioned in order to serve the HO requests available in the voice and data queues. The performance measures were evaluated with two types of scheduling and with different values of guard channels. We found that with QLT, there is a great improvement of data performance without inducing a perceptible degradation of the voice QoS. We also showed that QLT performs better when C_h is smaller and C_v is greater. Thus QLT combined with RCS and with the queuing strategy may improve the overall performance.

The comparison between analytical and simulation results showed that the proposed analytical model is fairly accurate. The simulation model has been generalized by considering a realistic scenario. Different bandwidth requirements were applied to voice and data users and burst allocation was considered for web users. Results showed that the QLT performance is still interesting with a generalized data model.

In the next chapter, we will extend our CAC model by integrating *adaptive* multimedia applications in order to accommodate multiple users with different bandwidth requirements. To this end, we will present a dynamic adaptive architecture that enhances the proposed CAC by coping with user's mobility and by dynamically adapting the mobile calls' bandwidth.

Chapter 5

Dynamic Adaptive Architecture: DYNAA

5.1 Introduction

In the previous chapter we proposed a Call Admission Control model that serves four classes of service. Two different queuing disciplines were introduced to enhance the model's performance. We used an analytical approach to evaluate our model. Moreover, we carried out a simulation to validate the model and to generalize the adopted assumptions.

In end-to-end QoS frameworks for multimedia wireless mobile systems, the high level of fluctuation in the availability of network resources is the major issue to be addressed. In order to provide quality of service to mobile flows in the presence of the scarce variable resources and user mobility, we propose to enhance our model.

In this chapter, we describe the development of a *DYNamic Adaptive Architecture* DYNAA that takes into account the user's mobility and adapts dynamically the bandwidth granted to the users.

We start this chapter by describing the challenges that must be addressed in wireless multimedia networks. In the subsequent section, we present our work's context. We then introduce our proposed architecture and discuss the different integrated layers. The proposed call admission control and bandwidth adaptation algorithms are described.

We also present enhanced scenarios of DYNAA together with their simulation results. A comparative evaluation of the studied architecture finalizes the study before concluding the chapter.

5.2 Challenges in Wireless Multimedia Environments

With the emergence of broadband wireless networks and the increasing demand of multimedia information on the Internet, wireless multimedia services are expected to become widely deployed. QoS provisioning in wireless networks presents a number of technical challenges that are discussed in the following paragraphs.

First, the bandwidth resource is limited in a wireless environment. The narrow bandwidth of a wireless network is a bottleneck, especially in the presence of high demanding bandwidth multimedia applications.

Secondly, the fluctuation in resource availability in wireless and mobile networks is much more severe than the level of fluctuation in fixed networks and this is due to different reasons:

1. The throughput of a wireless channel is often reduced due to multi-path fading, co-channel interference and noise.

2. The second main reason for the fluctuation in availability of network resources is mobility and handover. As a mobile station roams in the wireless network and is handed over from one access point to another, there is a change in the wireless and wired resources. This change in resources becomes critical when the terminal moves between different networks (e.g. from a wireless local area network to a wireless wide area network, the available bandwidth may vary drastically from a few Mega bits per second to a few Kilo bits per second).

There is a growing consensus that an adaptive quality of service model is a way of addressing these problems. Three techniques are to be studied:

1. Multimedia applications need to be adaptive and to accept the different QoS levels imposed by the network.

2. End systems must be network aware as they must be able to take the network status into account and adapt the multimedia application accordingly.

3. Networks must provide scalable application and must handle the adaptive QoS required by these applications. This introduces more requirements on the transport layer, routing, wireless medium, CAC, proactive resource reservation and signaling aspects. In our study, we are interested in the CAC implementation.

As a result, the end-to-end QoS provisioning is no longer the sole responsibility of the network nor of the application. Now, this responsibility is shared between the application and the network. They must together choose the QoS level to deliver multimedia content to a wireless mobile terminal in the most acceptable form given the available resources in the network.

Our focus is on developing a comprehensive framework that takes into account the above problems for wireless mobile networks which support multimedia applications. We propose a dynamic adaptive architecture DYNAA. The proposed DYNAA establishes a collaboration between the application and the network. This collaboration handles the user's mobility and the high variability in network conditions while offering the best possible service to the user.

5.3 Ambience Framework

The proposed architecture was conceived within the Information Technology for European Advancement (ITEA) AMBIENCE project. Ambient Intelligence is an exciting new paradigm in information technology, in which people are empowered through a digital environment that is aware of their presence and context, and is sensitive, adaptive, and responsive to their needs, habits, gestures and motions [ITE]. It is the merger of two important visions and trends: "ubiquitous computing" and "social user interfaces". It builds on advanced networking technologies, which allow robust, ad-hoc networks to be formed by a broad range of mobile devices.

By adding adaptive user-system interaction methods, based on new insights in the way people like to interact with computing devices (social user interfaces), digital environments can be created which improve the quality of life of people by acting on their behalf. These context aware systems combine ubiquitous information, communication, and entertainment with enhanced personalization, natural interaction and intelligence.

Ambient Intelligent environments can be characterized by important basic elements among others ubiquity, awareness, intelligence, and natural interaction. Ubiquity refers to a situation in which we are surrounded by a multitude of interconnected embedded systems, which are invisible and lie into the background of our environment. Awareness refers to the ability of the system to locate and recognize objects and people, and their intentions. Intelligence refers to the fact that the digital surrounding is able to analyze the context, learn from people behavior and adapt itself to the people who live in it.

In future networks, the intelligence should be everywhere: within the network elements and within the terminal. Consequently, the terminal and the network must cooperate in order to meet the QoS requirements of the users. This reveals the importance of the adaptation framework.

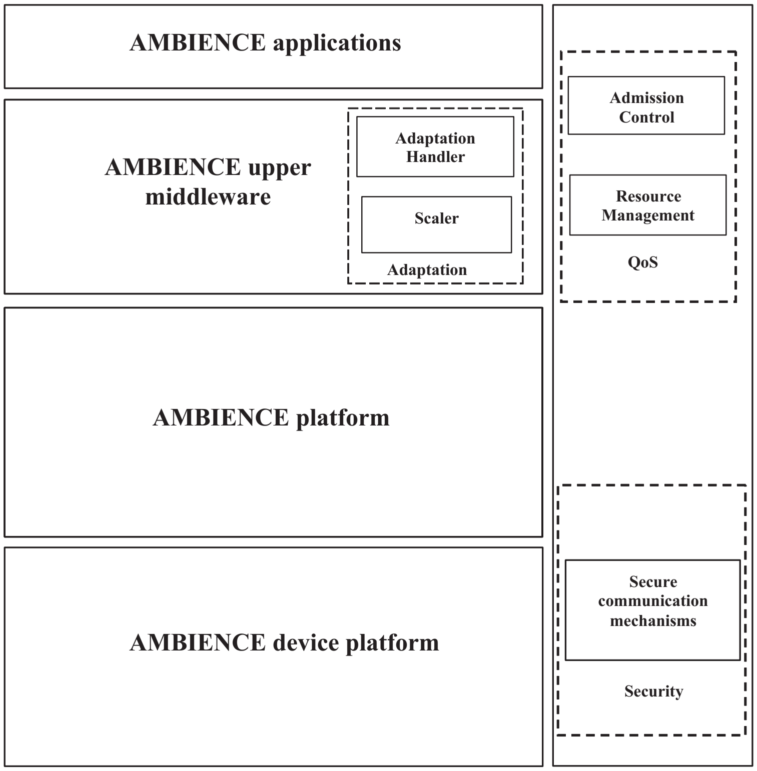


Figure 5.1: Ambience Global Architecture

Figure 5.1 illustrates the Ambience global architecture. This architecture is decomposed using a layered reference model. The upper part of the architecture provides the software environment specific to each type of application. This software is based on the top of the Ambience upper middleware which allows the users to access to the system's intelligence. The middleware controls the functional Ambience platform which is based in its turn on the top of the device platform. The quality of service and security functions operate at the different layers of the architecture and are represented here as a vertical layer.

In an attempt to improve the QoS, we propose a CAC within an innovative adaptive architecture DYNAA. As shown in figure 5.1, the second layer contains the adaptation module that is a part of the proposed DYNAA. The scaler and the adaptation handler are submodules of the adaptation module and are explained in the following section.

5.4 DYNAA

Figure 5.2 shows the dynamic adaptive architecture DYNAA that we propose. In our study, the focus is on the application and on the network adaptive layer. There is an interaction between these two layers that enable them to cooperate as shown in figure 5.2.

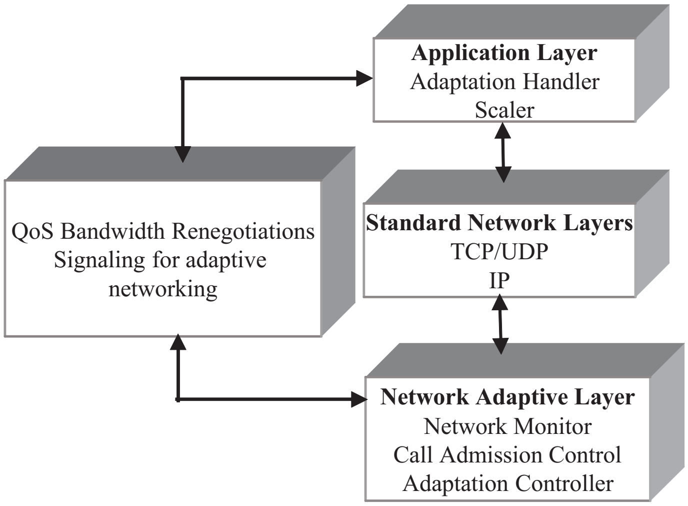


Figure 5.2: Dynamic Adaptive Architecture: DYNAA

5.4.1 Application Layer

The QoS components and functions that were designed for wired networks need to be enhanced and sometimes redesigned to provide multimedia services in mobile wireless networks.

Applications in wireless environments are expected to handle all resource fluctuations within acceptable bounds. As a network can offer to an application any QoS level within the predefined bounds, it can handle mobility, channel errors and dynamic resource availability by varying the allocation of resources. In order to be *adaptive*, multimedia applications must adapt to resource fluctuations. Thus the multimedia streams can adopt various coding approaches (transcoding, layered coding,...). With the layered coding approach, the multimedia stream is presented to the network in the form of a hierarchy of substreams or layers. These layers contain a base layer which must be always transmitted, and enhancement layers which refine the quality of the image. Depending on the resource availability, a subset of these substreams or layers is selected and transported by the network to achieve the best representation quality at a wireless mobile terminal.

If a cell is overloaded, the receiver only receives a subset of the multimedia substreams according to the decisions taken by the CAC and by the bandwidth adaptation algorithms. Otherwise, the receiver will receive the whole multimedia stream.

The application layer consists of two important modules: the adaptation handler and the scaler.

The adaptation handler informs the multimedia scaler about the allocated bandwidth, depending on the network decision. The scaler performs the adaptation: it distinguishes the multimedia substreams and drops the layers according to their significance. The scaler has the following advantages:

1. Reduction of the QoS degradation: the scaler understands the structure of multimedia streams and can thus selectively drops layers according to their significance. This is an improvement on random dropping which corrupts the multimedia stream. When faced with insufficient resources, the scaler starts by dropping the highest enhancement layer and continues down to the base layer.

2. Lower new call blocking and lower handover blocking probability: a request from a non adaptive application is often rejected since the required bandwidth is larger than the available bandwidth. In contrast, a request from an adaptive application is accepted as it can dynamically adapt to the available bandwidth.

5.4.2 Network Adaptive Layer

The network adaptive layer consists of specialized modules that support multimedia requirements. These modules are *the network monitor, the call admission control and the adaptation controller*. In order to provide adaptive service, the following sequence of events occurs (figure 5.3):

1. The network monitor periodically computes relevant QoS measures. These measures are the degree and the ratio of degradation and the current network load. We define how these measures are computed in section 5.4.5. They are used by the CAC when accepting new and HO calls.

2. The application notifies the network about its demand to set-up a call between end-points.

3. At this point, the network performs the call admission control using the measures computed by the network monitor. The CAC must ensure that even if it admits the mobile station, the QoS supplied to existing calls is maintained to a predefined level. In order to achieve this, the CAC and the resource allocation functions must take mobility into account. In other words, the CAC must be influenced by the changing network load created by mobility.

4. When performing the call admission control, an incoming call may cause a resource conflict between competing calls. Thus, the CAC interacts with the adaptation controller that adapts the resource allocation among the existing flows. The adaptation controller communicates with the bandwidth adaptation algorithms in order to resolve resource conflicts and to distribute available bandwidth among competing calls. The centralized adaptation controllers are located at the access points of the wireless network.

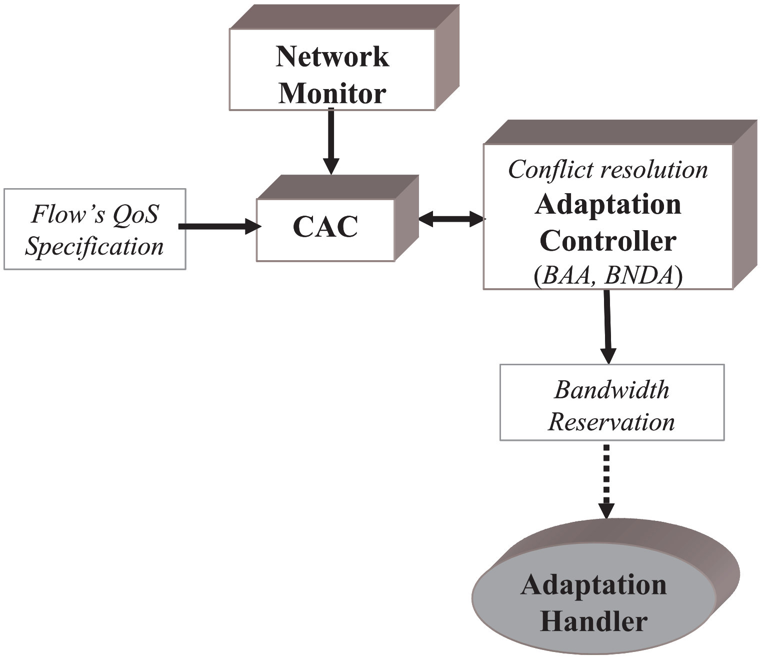


Figure 5.3: Network Adaptive Layer

5. DYNAA uses a centralized adaptation controller and distributed adaptation handlers in each cell. The adaptation controller notifies, through signaling, the adaptation handler about the bandwidth allocated. The adaptation handlers are responsible for enforcing the application to the adaptation process.

   The adaptation handlers determine whether or not the application will adapt to the available bandwidth. Typically, the adaptation controller allocates bandwidth to a flow and the adaptation handler decides whether to accept or not this allocated bandwidth.

5.4.3 Signaling Interface

A signaling interface provides various signaling features, flow set-up and adaptation control, between the mobile device and the access point. This interface supports signaling between the adaptation controller which determines the bandwidth allocations and the distributed adaptation handlers that are periodically informed of the bandwidth allocation.

5.4.4 Call Admission Control and Adaptation Algorithms: the Proposal

5.4.4.1 In the Literature

For adaptive multimedia applications, the existing QoS parameters for non-adaptive QoS parameters, such as the forced termination probability (or the call dropping probability), become trivial to be guaranteed at the expense of bandwidth degradation caused by adaptation. *It is to be noted that degradation is reached whenever the assigned bandwidth to a call is less than the requested bandwidth.*

A new QoS parameter, the Degradation Period Ratio DPR, is proposed by Kwon et al. in [KCBN99].

It represents the portion of a call's lifetime during which the call is degraded. The CAC proposed by the authors monitors the state of the cellular system at regular intervals and performs call admission in a distributed manner. With this CAC, the time average of the call degradation ratio in the given and the neighboring cells is measured. If this monitored average is inferior to a certain threshold, the call is accepted. Otherwise, it is rejected.

The new QoS parameter DPR does not characterize the degree of degradation. In order to fully characterize the bandwidth degradation and to provide better QoS to users, the authors in [XPCW01], [XPCW00] and [XPC01] propose two novel QoS parameters: the degradation ratio DR and the degradation degree DD. These two QoS parameters characterize both the frequency of degradation and the degree of degradation. The smaller the values of these parameters is, the better QoS.

In [XPCW01], the authors apply two types of fairness algorithms: the intra-fairness and the inter-fairness algorithms among the adaptive classes of service. Each class of users obtains an amount of bandwidth proportional to the class arrival rate. Thus, the performance obtained suffers from an under-utilization of the bandwidth.

Two orthogonal parameters are identified in [DSAB97] namely, the total carried traffic and the bandwidth degradation, such that the improvement of either of them leads to the degradation of the other. The authors formulate a "cost function" that describes the total revenue earned by the system from bandwidth degradation and call admission policies. Methods to compute the optimal policy maximizing the revenue is discussed.

In [KKC00], the authors choose the CAC thresholds in order to maximize the revenue. They take into consideration QoS parameters such as the call blocking probability and the call degradation probability.

The proposed scheme in [OKS98] allocates a bandwidth to a connection in the cell where the request originates and reserves bandwidth in all neighboring cells. The amount of bandwidth to reserve is dynamically adjusted, reflecting the current network conditions.

5.4.4.2 DYNAA Proposal

In the above-mentioned papers, the HO dropping probability becomes trivial to be guaranteed at the expense of the calls' degradation. Whenever a HO call requests service and if there are not available resources, the network adapts and even degrades the existing calls' bandwidth in order to satisfy the incoming call.

It is true that the forced termination of a call is a very frustrating phenomenon that may happen to a user. However, calls adaptation can be very annoying especially when it occurs frequently and when the adaptation of the calls' bandwidth leads to the calls' degradation. When a call requests a bandwidth and the network assigns a bandwidth less than the requested bandwidth to the call, the call is degraded. Thus, the adaptation framework must be carefully studied.

An efficient CAC in an adaptive framework must maintain a balance between two conflicting requirements: *minimize the handover call dropping probability* P_{drop} *and prevent the frequent adaptation*. In order to maintain an acceptable handover call dropping probability, the maximum number of calls should be adapted or even degraded into a network: this may result in unacceptably high degradation degree and degradation ratio due to insufficient resources for handover calls. Therefore, it is very important to have a *dynamic* CAC which can estimate future resource demands and degrade an acceptable number of calls to maintain an acceptable handover call dropping probability.

One of the critical tasks of a mobile computing environment is to prevent frequent adaptation due to the dynamics of resource and mobility of flows while optimizing the network performance (i.e. while having P_{drop} less than the required P_{drop}. The required P_{drop} will be defined by $P_{drop,qos}$). When the frequent adaptation is prevented, less signaling is generated on the radio bandwidth which constitutes a bottleneck in wireless networks.

Another important task is to take into account the current load when adapting the calls' bandwidth. Even, if the HO load decreases, the CAC in [XPC01] always accepts the HO request and that leads to the degradation of other calls in order to satisfy the incoming request. Consequently, the question that might be asked is as follows:

Why don't we dynamically adapt the calls' bandwidth while taking into account the current network load?

The above-mentioned tasks are achieved with DYNAA. Our proposed architecture dynamically adapts the calls' bandwidth based on the current network load. Moreover, the implemented algorithms within this architecture try to satisfy the QoS requirements of the classes of service. This is done by monitoring the average dropping probability and the degradation parameters and by adjusting the adapted bandwidth accordingly.

In the following subsections, we explain in detail the Call Admission Control algorithm, together with the Bandwidth Adaptation Algorithm *(BAA)* and the Bandwidth Adaptation with no Degradation Algorithm *(BNDA)*.

5.4.5 CAC Scheme

Before describing the proposed CAC, let us present the adopted classes of service in our study.

5.4.5.1 Classes of Service

Assume that a call is set-up between a sender and a receiver. When a sender transmits a multimedia stream to a receiver, the receiver will receive the whole multimedia stream or a subset of layers depending on the network decision. The receiver expects to receive at least a "requested" number of layers within the multimedia stream. Thus the network must allocate a "requested" bandwidth to the call. These "requested" number of layers and "requested" call' s bandwidth are specified by the user (receiver) when signing the contract. When the network allocates a bandwidth less than the requested bandwidth to a call, the call (or the application) is then degraded.

The requested call's bandwidth corresponds to an acceptable level of quality of service. However, ideally, the user would like to have a call with a bandwidth greater than the requested bandwidth. The bandwidth allocated to a call may vary within a specific range depending on the application's tolerance to degradation.

Some applications may tolerate that the network allocates to the calls a bandwidth much lower than that requested. Others cannot function correctly with this situation. In our study, we consider two classes of service:

- The Hard Adaptive class (HA): this class handles the applications that are not tolerant to frequent degradation: the applications using HA class require stringent guarantees and are more susceptible to QoS fluctuations. These applications permit to adapt the call's bandwidth with stringent constraints. The upper-bound values of DD and DR are relatively small.

- The Soft Adaptive class (SA): this class handles the applications that are more tolerant to degradation than the applications using HA class. Applications handled by SA class are less susceptible to the QoS fluctuations than the applications using HA class. They adapt the call's bandwidth with soft constraints. The upper-bound values of DD and DR are higher than those for HA class.

Subsequently, the HA and SA classes will be referred to as class 1 and class 2 respectively in order to respect the mathematical notation that will be defined in the next section.

5.4.5.2 Resource Specification

For an application using a service class i (HA or SA class), a call requests a bandwidth denoted by $b_{i,req}$. The requested bandwidth $b_{i,req}$ must vary within the range $\{b_{i,1}, b_{i,2}, \ldots, b_{i,j}, \ldots, b_{i,k_i}\}$ where:

- k_i is the number of multimedia layers of the application handled by the class i of service.

- $b_{i,j}$ is the bandwidth of j layers of an application handled by the class i of service, such that $b_{i,j} < b_{i,j+1}$ for $j = 1, 2, \ldots, k_i - 1$.

- $b_{i,1}$ is the minimum bandwidth, b_{i,k_i} is the maximum bandwidth that can be allocated to the call related to the application that is handled by the service class i.

- $b_{i,req} > b_{i,1}$. We assume that the calls related to applications that are handled by the same class of service use the same requested bandwidth.

An example of this adaptive multimedia application is a video stream (A) encoded with *MPEG-2 (I)* frames, *MPEG-2 (I+P)* frames and *MPEG-2 (I+B+P)* frames (figure 5.4). In this example, we have one class of service (i.e. $i = 1$). k_1 equals 3, b_1 is the bandwidth needed to transmit *MPEG-2 (I)*, b_2 is the bandwidth needed for *MPEG-2 (I+P)* and b_3 is the bandwidth needed for *MPEG-2 (I+B+P)*.

Let us assume that the source is transmitting the same video stream to three users (X,Y,Z) located respectively in the same wireless LAN as the source, in a cellular network and in another network separated by the Internet. As discussed before, in a network providing a suitable call admission and bandwidth adaptation algorithms, one can avoid unnecessary call blocking if existing calls accept lower quality of service level based on their contract with the network. In this example, users Y and Z are switched to a lower QoS level because of a lack of resources. Hence, Y may just receive *MPEG-2 (I)* frame and Z the *MPEG-2 (I+P)* frames. As for X, the wireless LAN can allocate sufficient resources for *MPEG-2 (I+B+P)* frames.

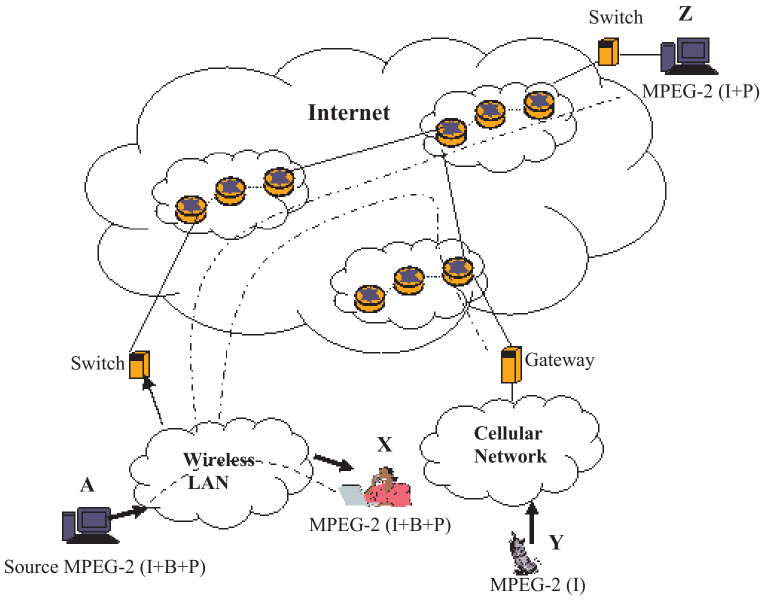


Figure 5.4: Adaptive Application

5.4.5.3 CAC Algorithm

As shown in section 5.4.2, the network monitor computes periodically the degradation parameters for each class of service and the dropping probability. These measures are passed afterwards to the CAC module. The degradation ratio and the degradation degree of the class i of service are denoted by DR_i and DD_i respectively. We refer to $DD_{i,qos}$ and $DR_{i,qos}$ as the upper-bound values of the degradation parameters of a class i.

At this point, let us give the definitions of some parameters used in the model:

- $x_i(t)$ stands for the number of calls of class i in a cell at time t.
- $b_{i,assi}(s,t)$ denotes the assigned bandwidth for a call s of a class i at time t. To be more precise, $b_{i,assi}(s,t)$ belongs to the range $\{b_{i,1}, b_{i,2}, \ldots, b_{i,j}, \ldots, b_{i,k_i}\}$ such that $1 \leq s \leq x_i(t)$. Note that if $b_{i,assi}(s,t) < b_{i,req}$ then the call s is degraded.
- $I(f)$ is the indicator function which returns 1 if f is true and 0 otherwise.

If n is the number of classes of service (equal to 2 in our study), ΔT a time interval for the measurement, and τ a time variable, the degradation parameters are such that [XPCW01]:

$$For\, i = 1, \ldots, n$$

$$DR_i(\tau) = \frac{1}{\Delta T} \times \int_{\tau-\Delta T}^{\tau} \frac{\sum_{k=1}^{x_i(t)} I(b_{i,assi}(k,t) < b_{i,req})}{x_i(t)} \times dt \tag{5.1}$$

```
New Call Arrival:
If $((DD_i \leq DD_{i,qos})$and $(DR_i \leq DR_{i,qos}))$ for $i = 1, .., n$ then
accept the new call and call the BAA procedure
else reject the call

Handover Arrival:
If $(P_{drop} \geq P_{drop\_max})$ then
accept the call and call the $BAA$ procedure.

If $(P_{drop} \leq P_{drop\_min}$ ) then
If (( $DD_i \leq DD_{i,qos}$ )and $(DR_i \leq DR_{i,qos}$ )) for $i = 1, .., n$ then
accept the HO and call the $BNDA$ procedure
Else reject the HO

If $(P_{drop\_min} < P_{drop} < P_{drop\_max})$ then
If $((DD_i \leq DD_{i,qos})$and $(DR_i \leq DR_{i,qos}))$ for $i = 1, .., n$ then
accept the HO and call the $BAA$ procedure
Else reject the HO
```

Table 5.1: CAC Algorithm

$$DD_i(\tau) = \frac{1}{\Delta T} \times \int_{\tau-\Delta T}^{\tau} \frac{\sum_{k=1}^{x_i(t)} I(b_{i,assi}(k,t) < b_{i,req})(b_{i,req} - b_{i,assi}(k,t))}{(b_{i,req} - b_{i,1}) \sum_{k=1}^{x_i(t)} I(b_{i,assi}(k,t) < b_{i,req})} \times dt \tag{5.2}$$

One can see that these equations compute the time average of the degree of degradation and the frequency of degradation. Since the events of changing $x_i(t)$ happen discretely in time, the above integrations become discrete sum when implementing them. Both DR_i and DD_i take value ranging from 0 to 1. The smaller the values of the degradation parameters is, the better QoS.

In our CAC algorithm (table 5.1), a new call is accepted only if the degradation parameters are less than the corresponding upper-bound values $DD_{i,qos}$ and $DR_{i,qos}$.

For HO calls, we introduce two thresholds for the call dropping probability, namely P_{drop_min} and P_{drop_max}. If P_{drop} is greater than P_{drop_max}, then the HO load is relatively high. Hence, the HO call is accepted without testing the degradation parameters and the Bandwidth Adaptation Algorithm *(BAA)* is executed (see subsection 5.4.6).

On the other hand, if the monitored measure P_{drop} is less than P_{drop_min} (respectively between P_{drop_min} and P_{drop_max}), then the HO load is relatively low. Thus, we accept the call if the performance degradation parameters are less than upper-bound values. Afterwards, *BNDA* (respectively *BAA*) is executed.

Note that the threshold P_{drop_max} represents a fraction of the required $P_{drop,qos}$. It was introduced in order to better control the value of P_{drop} and in order to keep P_{drop} less than

$P_{drop,qos}$.
In our work, the proposed CAC aims to have a compromise between P_{drop} and the degradation parameters where the adaptation proposed is a function of the mobility through measuring P_{drop}.

The acceptance by the CAC does not guarantee the call's request to be finally accepted. It also depends on whether or not the bandwidth adaptation algorithm can allocate enough bandwidth for the "accepted" call. The adaptation algorithms are explained in the next subsections.

```
Call Arrival of class $i$
If ($A \geq b_{i,req}$ ) then
    Assign_at_most($b_{i,max}$)
Else
{
    Reduce($SA$,$b_{2,req}$);
    If ($A < b_{i,req}$ ) then Reduce($HA$,$b_{1,req}$)
    If (($A \geq b_{i,req}$ ) or (new call arrival)) then
        Assign_at_most($b_{i,max}$)
    Else //$A < b_{req}$ and the call is a HO request
    {
        Reduce ($SA$,$b_{2,min}$)
        If ($A < b_{i,min}$) then Reduce ($HA$,$b_{1,min}$)
        If ($A < b_{i,min}$) reject the call
        Else Assign_at_most($b_{i,req}$)
    }
}
Reduce (class, b)
{
    While (($A < b_{i,req}$) and (there exists a call in class
    such that its bandwidth is $> b$))
    {find the largest bandwidth in class
    reduce its bandwidth to $b$}
}
Assign_at_most (level)
{ Assign a bandwidth as much as possible but at most level.}
$A = available\ bandwidth$
```

Table 5.2: Bandwidth Adaptation Algorithm *(BAA)*

5.4.6 Bandwidth Adaptation Algorithm/ Bandwidth Adaptation with No Degradation Algorithm

The Bandwidth Adaptation Algorithm (*BAA*) and the Bandwidth Adaptation with No Degradation Algorithm (*BNDA*) decide the changes in the calls' bandwidth adaptively when there is a call arrival.

```
**Call Arrival of class $i$**
If ($A \geq b_{i,req}$) then Assign at most($b_{i,max}$)
Else
{
Reduce($SA$,$b_{2,req}$)
If ($A < b_{i,req}$) then Reduce($HA$,$b_{1,req}$)
If (($A \geq b_{i,req}$) or (new call arrival)) then
Assign_at_ most($b_{i,max}$)

Else{ // $A < breq$ and the call is a HO request
If ($A \geq b_{i,min}$) Assign at most($b_{i,req}$)
Else reject the call } }
```

Table 5.3: Bandwidth Adaptation With No Degradation Algorithm *(BNDA)* Algorithm

Ideally, every call in a cell must be allocated the maximum bandwidth. However, if the cell is over-loaded, some of the calls may receive a bandwidth which is lower than their requested bandwidth. If a new or handover call arrives, some of the calls already in progress in the cell may be forced to lower their bandwidth to accommodate the newly arrived call. On the other hand, if a call terminates or hands over to another cell, some of the remaining calls in the cell can readjust and increase their bandwidth. The proposed bandwidth adaptation algorithms use the above ideas to allocate, increase and decrease calls' bandwidth.

In order to prioritize Hard Adaptive HA class, the adaptation algorithm is performed over the Soft Adaptive SA class first, and then over the HA class. Within the same class, the adaptation algorithm is first applied to the least degraded calls.

For the reader convenience, we denote the minimum bandwidth of a user in a class i of service $b_{i,1}$ by $b_{i,min}$, the maximum bandwidth of a user in a class i of service b_{i,k_i} by $b_{i,max}$.

When a call belonging to a class i requests to be served with a $b_{i,req}$ bandwidth, the *BAA* and *BNDA* algorithms behave as follows.

With *BAA* (table 5.2) , if the available bandwidth (A) is less than $b_{i,req}$, then our algorithm tries to lower the bandwidth of calls belonging to SA class to their requested bandwidth $b_{2,req}$. The same procedure is executed afterwards over calls of HA class.

If A is still less than $b_{i,req}$, then the call is rejected if it is a new call. In case of a HO call then the calls' degradation is performed: *BAA* tries to squeeze, to the minimum bandwidth $b_{2,min}$, the bandwidth of calls belonging to SA class and then that of calls of HA class.

As for *BNDA*, some adaptation is made without call's degradation (table 5.3). *BNDA* behaves as *BAA* but differs in the HO handling. If after reducing the calls bandwidth to their requested bandwidth A is still less than $b_{i,req}$, then no calls' degradation is performed and the HO is blocked.

When there are some free resources due to a call departure (at the end of a call or after a HO), the new available bandwidth must be used in order to satisfy the degraded existing calls.

Table 5.4 describes the bandwidth reallocation algorithm whenever there is a departure. Since the available bandwidth is increased, some calls in the cell can be upgraded. First, we pick up the most degraded call in the HA class and we increase its bandwidth to the requested bandwidth. We keep doing this until there is no available bandwidth or until every call in the HA class has a bandwidth larger than or equal to the HA's requested bandwidth $b_{1,req}$. Then, the same procedure is applied to SA class if the available bandwidth is still greater than zero. Afterwards, we try to increase the bandwidth of the call with the smallest bandwidth in HA class. The same thing is done then to SA class if the available bandwidth is still greater than zero.

```
**Call Departure**
While (($A > 0$)and (exist degraded calls in class $HA$))
{Find the most degraded call in $HA$
Assign_at_most($b_{1,req}$)
}
While (($A > 0$)and (exist degraded calls in class $SA$))
{Find the most degraded call in $SA$
Assign_at_most($b_{2,req}$)
}
While ($A > 0$)
{Find the smallest bandwidth in $HA$ class
Assign_at_most($b_{1,max}$)
}
While ($A > 0$)
{Find the most degraded call in $SA$
Assign_at_most($b_{2,max}$)
}
Assign_at_most ($level$)
{Assign a bandwidth as much as possible but at most $level$.}
$A$ =available bandwidth
```

Table 5.4: Departure Procedure

5.5 Simulation

5.5.1 Scenarios

In order to evaluate the performance of the proposed architecture, some simulations were carried out with different scenarios.

1. The first scenario implements DYNAA as proposed in the previous section. In this scenario (referred to as DYNAA), HO requests are blocked when they are refused by the CAC and the bandwidth adaptation algorithms. As a result, it is a blocking scheme for both new and HO requests.

2. The second proposed scenario (referred to as DYNAA_Wait) implements DYNAA architecture with a waiting alternative. In this case, the Network Adaptive Layer is depicted in figure 5.5.

   In this scenario, if after applying the BAA and $BNDA$ algorithms, there are still no available resources, the HO requests are not rejected. Instead, the HO requests belonging to SA and HA class (SA_HO and HA_HO requests) are inserted in two distinct queues.

   In order to schedule the SA_HO and HA_HO requests, the Queue Length Threshold *(QLT)* scheduling policy is used: QLT gives priority to SA calls whenever the number of queued SA_HO requests is above some threshold (Lth). Otherwise, the HA_HO requests are served first. The performance of QLT in a mobile system was tested in the previous chapter using an analytical and a simulation approach, and we have proven that QLT may improve the performance of a mobile system.

   A queued HA_HO request is deleted from the queue when it passes through the HO area before getting a channel or if its communication is completed before passing through the HO area.

   SA connections typically are more tolerant to delay as compared to HA connections. Thus, whenever a SA request is not satisfied within the current cell, it is transferred to the SA queue of the target cell. Recall that deleting the request might impose a penalty incurred to call re-initiation and data retransmission.

   Note that the call departure procedure differs in this scenario because of the waiting HO requests. In fact, whenever there are some free resources due to a call departure, the new available bandwidth must be used in order to satisfy the HO requests waiting in the queues first and then the degraded existing calls.

   Therefore, we apply the call admission control algorithm which tests P_{drop} and the degradation parameters as precised in subsection 5.4.5.3. Afterwards, the scheduler interacts with the adaptation controller and applies the adaptation algorithm according to the decision of the CAC.

   If after serving the HO requests, there are still some available resources, existing calls are upgraded as precised in Table 5.4. The call departure procedure with DYNAA_Wait can be summarized in the following set of rules:

   - If $P_{drop} \geq P_{drop\_max}$, the HO requests are served and scheduled according to QLT. In this case, the BAA is executed. If there is still some available bandwidth after serving the HO requests, the unused bandwidth helps to upgrade existing calls.
   - If $P_{drop} \leq P_{drop\_min}$, we test if the degradation performance parameters are respected. If this is the case, we allocate available bandwidth to the waiting HO requests after applying the $BNDA$ algorithm. Otherwise, the existing calls are upgraded.
   - If $P_{drop\_min} < P_{drop} < P_{drop\_max}$, we test if the degradation performance parameters are respected. If this is the case, we allocate available bandwidth to the waiting HO requests after applying the BAA algorithm. Otherwise, existing calls are upgraded.

3. The third simulated scenario is a static one. In fact, it implements the proposed architecture DYNAA but it is static as it does not adjust the calls' bandwidth to the HO load. Whenever

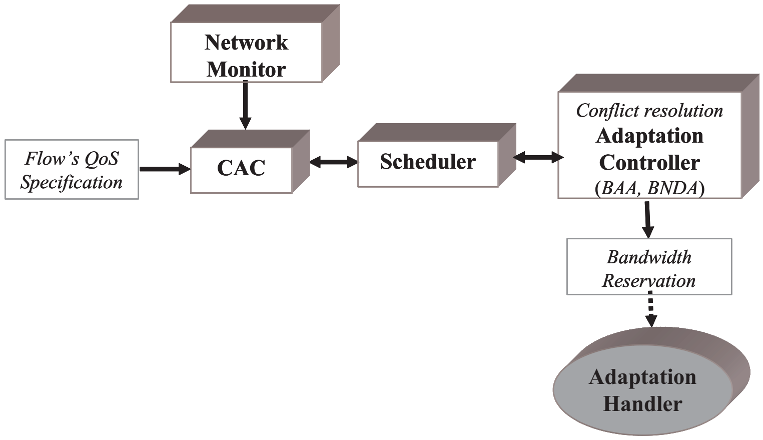


Figure 5.5: Network Adaptive Layer with DYNAA_wait

there is a HO request, we always accept the call without checking neither the HO load nor the degradation parameters. This scenario is referred to as the static scenario.

4. The fourth simulated scenario does not support adaptation. This scenario is the one suggested in the previous chapter. More specifically, new SA (respectively HA) calls are blocked whenever the number of used channels is greater than $C - C_v$ (respectively $C - C_h$); where C is the total number of channels in the cell, C_h is the number of guard channels reserved for HO calls, C_v is the number of channels that are forbidden to new SA calls. In order to prioritize HA over SA calls, C_v is greater than C_h ($C_v = 12$, $C_h = 6$). The HA and SA handover requests can wait in two queues which are served with the QLT technique. In this scenario, we consider that the bandwidth requirement of HA calls is 3 channels, that of SA calls is 6.

5.5.2 Simulation Model

Computer simulations have been conducted by assuming a seven cell network. The edges of the simulated space wrap around to the opposite edges so that each cell has a complete set of interfering cells: thus, the border effect is avoided. The considered cells are assumed to have a radius of $1Km$, and a capacity of 60 channels. The users are vehicular with an average speed of $40Km/h$. The cell dwell time is exponentially distributed with mean $141.37s$, the dwell time in the HO area is exponentially distributed with mean $14s$. The length threshold L_{th} is equal to 2. Other parameters are depicted in table 5.5.

The HA traffic in the simulation is represented by an ON/OFF traffic with a rate of $12,2Kb/s$ within the ON period (chapter1). The arrival of a HA call in the simulation is assumed to be Poisson with mean rate λ_{nv}. Its duration is exponentially distributed with mean $120s$.

The SA traffic considered is represented by a typical WWW session (UDD 64 Kb/s) that consists of a sequence of packet calls. The parameters and the laws that model the web session are depicted in chapter 1. The arrival of a SA call in the simulation is assumed to be Poisson with mean rate λ_{nd}, such that $\lambda_{nv} = \lambda_{nvd} = \lambda_{ntot}/2$, where λ_{ntot} is the total call originating rate density.

In the simulation, we have considered applications that can adjust their coding rate and hence can adapt the assigned bandwidth to the offered network resources. Note that simulation results were retrieved with a 95% confidence interval (see appendix).

Parameter	**Value**
Number of Cell channels C	100
Class 1 (HA) ($b_{1,1}, b_{1,2}, b_{1,3}$)	$(1, 2, 3)$
Class 2 (SA) ($b_{2,1}, b_{2,2}, b_{2,3}$)	$(2, 4, 6)$
Requested Bandwidth ($b_{1,req}, b_{2,req}$)	$(2, 4)$
Measurement Time Interval ΔT	$5s$
Upper-bound values of Class 1 Degradation Parameters ($DR_{1,qos}, DD_{1,qos}$)	$(0.01, 0.01)$
Upper-bound values of Class 2 Degradation Parameters ($DR_{2,qos}, DD_{2,qos}$)	$(0.1, 0.1)$
P_{drop} Thresholds ($P_{drop_{min}}, P_{drop_{max}}$)	$(0.0075, 0.009)$
Required P_{drop} : $P_{drop,qos}$	0.01
Mean HA Call Duration	$120s$
Mean Cell Dwell Time	$141.57s$
Mean HO Area Dwell Time	$14s$

Table 5.5: Simulation Parameters

5.6 Numerical Results

5.6.1 Comparison between Static protocol and *DYNAA*

In order to evaluate the performance of DYNAA, let us compare the performance measures obtained with DYNAA (scenario 1) and with the static scenario (scenario 3).

The aim of this comparative study is to prove the utility of the dynamic adaptation proposition, that is to prevent frequent adaptation while still keeping the dropping probability less than a certain threshold. This is achieved by dynamically adapting the calls' bandwidth while taking into consideration the mobility of the flows. The exchange between the degradation parameters and the forced termination probability is illustrated hereafter.

Let X be a variable measured by simulation. We define ΔX as the relative variation of X measured with the static scenario and with DYNAA. Hence,

$$\Delta X = \frac{X_{static} - X_{DYNAA}}{X_{static}} \tag{5.3}$$

If $\Delta X > 0$ then the value of X is greater with static scenario than with DYNAA. If X is equal to DD or DR and if $\Delta X > 0$ then DYNAA behaves better than the static scenario. Recall that the smaller the values of the degradation parameters, the better the QoS.

Table 5.6 indicates that ΔDD is equal to 33.1% for HA class and to 23.4% for SA class. On the other hand, ΔDR attains the value of 32.3% for HA and 25.8% for SA class. DYNAA adjusts the adapted calls' bandwidth to the HO load. This explains why we obtain the degradation measures in DYNAA smaller than those for the static scenario.

The degradation parameters decrease is made at the expense of the increase of P_{drop} (figure 5.6). P_{drop} with DYNAA is greater than that of the static scenario where P_{drop} is almost

ΔDD_1	ΔDD_2	ΔDR_1	ΔDR_2
33.1%	23.4%	32.3%	25.8%

Table 5.6: Relative Variation of Degradation Performance between DYNAA and the static scenario

negligible. Nevertheless and as expected, P_{drop} obtained with DYNAA is less than the $P_{drop,qos}$ (1%). Consequently, the required performance is respected with the dynamic adaptive scenario. Note that the dropping probability increases until the call originating rate density of $0.5 calls/s/Km^2$, after which it maintains its stability at the value of 0.9%.

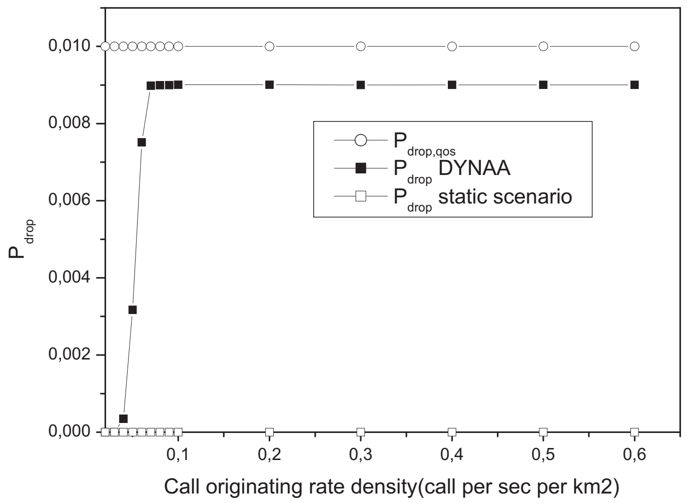


Figure 5.6: Call Dropping Probability Comparison between DYNAA and static scenario

5.6.2 Comparison between Non adaptive scenario and DYNAA_Wait

The next simulation results compare the performance of the enhanced scenario of DYNAA, DYNAA_Wait (scenario 2) to that of the non adaptive scenario (scenario 4). Simulation results show that the new call blocking probabilities of SA (B_{nSA}) and that of HA calls (B_{nHA}) with the non adaptive scenario are greater than those of DYNAA_Wait (figures 5.7, 5.8).

In fact, the adaptation with DYNAA_Wait permits to adapt the bandwidth of existing calls and to provide resources to new calls. It is to be noted that the decrease of the B_{nSA} is greater than that of B_{nHA}. This results from the choice of the values of the thresholds C_v and C_h. In fact, we block more new SA calls than new HA calls because the SA blocking threshold is less than for the HA blocking threshold.

As for the forced termination blocking probability (P_{drop}), figure 5.9 illustrates that with DYNAA_Wait, P_{drop} is always less than $P_{drop,qos}$. Indeed, in our implemented CAC and adaptive

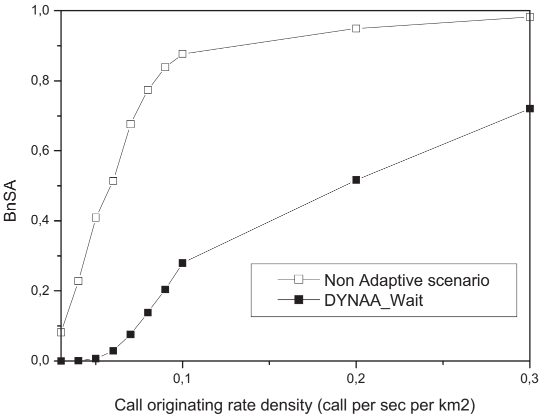


Figure 5.7: Blocking Probability of new SA calls (B_{nSA}) with DYNAA_Wait and with the non adaptive scenario

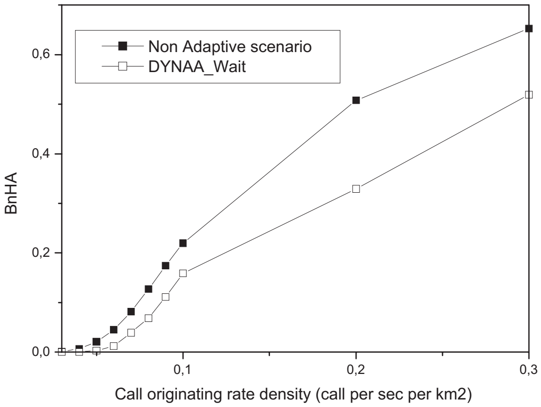


Figure 5.8: Blocking Probability of new HA calls (B_{nHA}) with DYNAA_Wait and with the non adaptive scenario

algorithms, we were concerned about maintaining this value less than P_{drop_max}. This was done in order to keep the P_{drop} value below $P_{drop,qos}$.

On the other hand, with the non adaptive scenario and with a call rate density less than $0.15 calls/Km^2/s$, P_{drop} is less than that of DYNAA_Wait. This results from the fact that priority is attributed to the HO calls in the non adaptive scenario by reserving some guard channels to the HO calls. Consequently, this may improve the dropping probability in non adaptive scenario and so the P_{drop} of the non adaptive scenario is smaller than that of DYNAA_Wait with a 95% confidence interval.

However, with the increase of the call rate density, P_{drop} in the non adaptive scenario increases. When it exceeds the value of $0.16\, call/s/Km^2$, it becomes greater than the required P_{drop}. On the contrary, DYNAA_Wait controls the P_{drop} and keeps it less than the required P_{drop}. This is done by dynamically adapting the calls' bandwidth.

As for the waiting delay encountered by the *SA_HO* requests, figure 5.10 reveals that the mean delay with DYNAA_Wait is less than that of the non adaptive scenario. This result is expected because the adaptive scenario serves more HO calls by adapting other calls. Thus, the waiting delay is reduced.

On the other hand, we have measured the mean HA waiting delay which is encountered by the *HA_HO* requests in the HO area. This measured delay is the mean delay of the *HA_HO* calls before being served by the neighbor cell or before the HO requests expired while being served by the current cell. For the SA requests however, it is a different matter. In fact, if the *SA_HO* requests are not served by the current cell, they will be transferred to the following cell. Consequently, the mean waiting delay T_{delay} is the mean delay of the *SA_HO* requests before being served by the neighbor cell or before being transferred to the waiting queue in the next cell.

As we can see in figure 5.11, the T_{delay} experienced by the HA class with the non adaptive scenario is less than that with the DYNAA scenario, for small values of the call originating

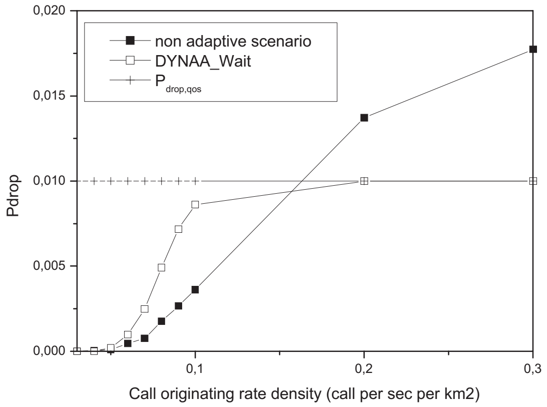


Figure 5.9: Call Dropping Probability with DYNAA_Wait and with the non adaptive scenario

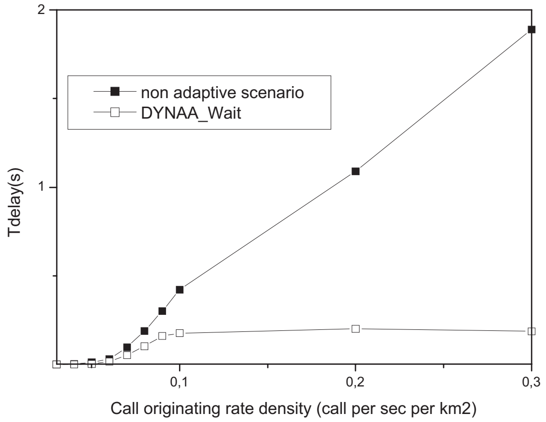


Figure 5.10: Mean SA Waiting time with DYNAA_Wait and with the non adaptive scenario

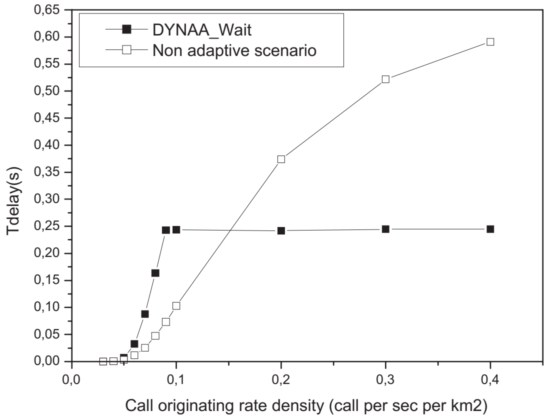


Figure 5.11: Mean HA Waiting time with DYNAA_Wait and with the non adaptive scenario

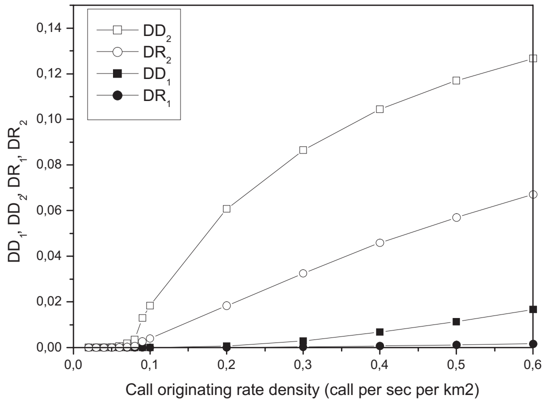


Figure 5.12: Degradation Parameters with DYNAA_Wait and with the non adaptive scenario

densities.

This result is obtained from the fact that priority is attributed to the HO calls in the non adaptive scenario by reserving some guard channels to the HO calls. Thus, this improves the delay of HA in the non adaptive scenario and consequently the T_{delay} of the non adaptive scenario may be smaller than that of DYNAA_Wait. However, for a call originating density greater than 0.15 $calls/Km^2/s$, the T_{delay} of the non adaptive scenario becomes greater than that of DYNAA_Wait and it continues to increase. Note that the delay of the HO request in the DYNAA_Wait becomes stable for a call originating density equal to 0.09 $calls/Km^2/s$.

It can be seen that the decrease of the new call blocking probabilities, of the dropping probabilities and of the delay with DYNAA_Wait is done at the expense of the bandwidth degradation. With the non adaptive scenario, the degradation parameters are zero because in this case calls do not suffer from degradation.

Figures 5.12 illustrates the fact that the degradation ratio of the HA and the SA class is respected over the entire range of call originating density. As for DD_1 (respectively DD_2), it is respected until $0.473 call/sec/Km2$ (respectively until $0.373 calls/sec/Km2$). Hence, by choosing the convenient rate density, one can limit the degradation below a certain threshold. At this point, we can conclude that DYNAA_Wait improves the overall performance at the expense of a slight controlled degradation.

5.7 Conclusions

The provision of QoS guarantees in wireless mobile networks is a complex problem. In this chapter, an admission control based on adaptive bandwidth algorithms is proposed in order to provide QoS to two classes of service.

It is shown that, by dynamically adjusting the allocated bandwidth while taking into account the current network conditions, our proposed architecture can achieve better QoS. With this dynamic adaptive approach, the frequent adaptation is prevented. This helps to reduce the signaling overhead.

The proposed architecture has been improved by permitting the HO requests to wait in the HO area before being served. The results obtained by simulations have illustrated the improved characteristics of the proposed architecture.

DYNAA_Wait can improve the overall performance at the expense of slight degradation which is controlled and limited. However, the system parameters have to be correctly chosen to optimize the degradation parameters, while maintaining the dropping probability stable over the range of the call originating rate density.

An interesting extension to our approach is the study of the signaling aspect between the network adaptive layer and the application layer. As a matter of a fact, the signaling is quite important to study because it carries the information between the user and the network. On the other hand, we are intending to study traffic models of higher burstiness than the models adopted in this chapter. The monitoring time interval ΔT will be adapted to the traffic models. A small time interval leads to a more precise application adaptation, but also increases the signaling load. Thus, there is a trade-off between the application adaptation, the frequency of the measurements and the signaling load. The fine tuning of the measurement time interval is an important task to be achieved by the network designer.

The promising results of our architecture lead us to integrate it within Cellular IP networks. Hence, we shall study the QoS aspect in chapter 7 by integrating DYNAA_Wait into a Cellular IP network. Before presenting this QoS aspect, we shall optimize first the uplink routing in Cellular IP networks. An optimized uplink routing study as well as a handover study will be detailed in the next chapter.

Chapter 6

Smooth Handover and Optimized Uplink Routing in Cellular IPv6 Networks

6.1 Introduction

Provision of various real time multimedia services to mobile users is the main objective of the next generation wireless networks, which will be IP-based and are expected to interwork with the Internet backbone seamlessly. Two major challenges exist in wireless mobile networks, namely the support of fast handover and the provision of Quality of Service over IP-based wireless access networks.

Research efforts are oriented towards the development of different micro-mobility protocols that can handle the IP mobility seamlessly. IP micro-mobility protocols are designed in order to limit the disruption to user traffic during handover and to handle frequent handovers across multiple subnetworks. IP micro-mobility protocols complement Mobile IP by providing fast and also seamless handover control.

In chapter 3, we presented the different IP micro-mobility protocols that are proposed for wireless mobile IP networks. Then, we exposed a comparative study of these micro-mobility protocols, taking into consideration major handover management issues. We noticed that each protocol has both strengths and weakness.

It is difficult to claim that one protocol is better than another one: the choice of a micro-mobility protocol depends on the priorities imposed by the network management.

In our work, we chose to study Cellular IPv6 because this micro-mobility protocol presents some important features (chapter 3), such as a cheap passive connectivity, an efficient location management, an efficient routing and a flexible handover.

Nevertheless, the Cellular IPv6 (CIP) protocol lacks two important issues: an optimized uplink routing and a QoS support. In this chapter, we address the uplink routing issues. In the next chapter, a QoS study is carried out.

The organization of the chapter is as follows. First, we study the uplink routing for intra-network traffic in Cellular IPv6 networks. We propose an enhancement for this mechanism.

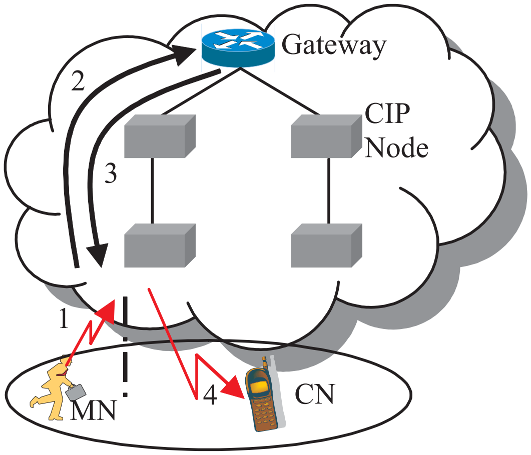


Figure 6.1: Non Optimization of Uplink Routing Mechanism for Intra-network Traffic in Cellular IPv6 Network

Secondly, a smooth and anticipated handover is presented. Simulations and performance analysis are presented before concluding the chapter.

6.2 Uplink Routing in Cellular IPv6 Networks

Regular data packets transmitted by mobile nodes are used to establish host location information in order to minimize control messaging. Uplink packets are routed from a mobile node to the gateway on a hop-by-hop basis. The path taken by these packets is cached in intermediate nodes.

Cellular IPv6 suffers from a non-optimization of the uplink routing in the case of intra-network traffic. The intra-network traffic is the traffic exchanged between two mobiles in the same Cellular IPv6 network. With Cellular IPv6, the traffic coming from the mobile node (MN) must pass through the gateway before being delivered to the corresponding node (CN), even if the MN and the CN are connected to the same base station. Far from an optimal path, this kind of routing increases the delay and the jitter of the packets. It also implies the waste of bandwidth which constitutes a problem in case of high traffic (figure 6.1).

D. Gatzounas et al. [SMG$^+$01] proposed an uplink routing optimization for Cellular IP networks that improves the protocol performance. In their proposition, even if the corresponding node and the mobile node are not in the same Cellular IP domain, all the Route caches in the route towards the gateway will be checked in order to find if CN has an entry with a “Tear down optimized flag” unset. This would cause an unnecessary delay for the cache processing. On the other hand, the overall performance is made at the expense of an increase of the signaling load.

In our study, we achieve two objectives. First, we enhance the uplink routing mechanism by adopting the mechanism proposed in [SMG$^+$01] and by improving it in order to have faster lookup, better scalability and less frequent signaling messages. Second, we propose to study the performance of a smooth and anticipated handover in Cellular IPv6 networks. More precisely, we evaluate the performance of a buffering mechanism while taking into consideration the optimized uplink routing mechanism.

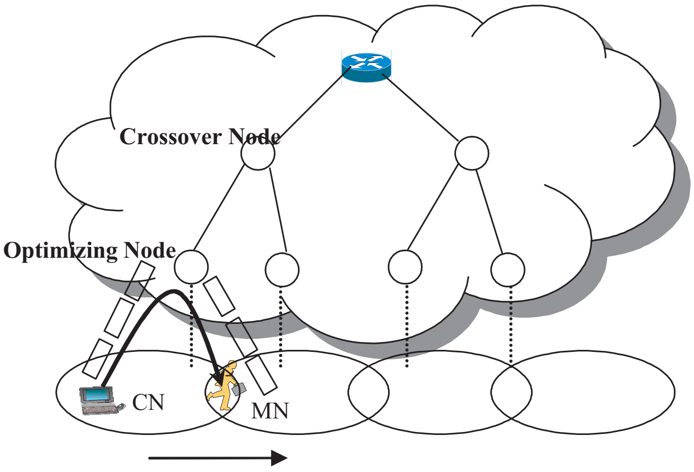


Figure 6.2: Optimizing and Crossover Nodes in Cellular IPv6 Networks

6.3 Protocol Extensions

Before detailing the optimized uplink routing mechanism, let us distinguish between the two following types of CIP nodes (figure 6.2):

1. *Optimizing Node.* This node is at the intersection of two paths: one path is from the MN's access node to the gateway router and the second path is from the CN's access node to the gateway router. Only one optimizing node at any time performs route optimization for a single pair of communicating mobile nodes in the same CIP network. The optimizing node must route data through the optimal path to the destination address.

2. *Crossover Node.* This node is at the intersection of two paths: one path is from the gateway to the previous access node (base station) and the second path is from the gateway to the new access node (base station). The crossover node has two mappings for the mobile during the handover.

Let us define the following messages and flags added to the Cellular IPv6 protocol:

1. *Proxy route-update* message: it is sent by the optimizing node towards the gateway [SMG$^+$01]. It is an IPv6 packet carrying a Hop-by-Hop Options extension header.

2. *Remove Mapping* message: it is sent by a mobile upon handover. It is an IPv6 packet carrying a Hop-by-Hop Options extension header.

3. *Optimize Route (OR)* flag : it is carried in the Hop-by-Hop Options extension header of the data packets. When set, the route optimization must be performed by the nodes receiving the packets (default=0).

4. *Duplicate (DUP)* flag: it is in the IP header of the data packets indicating, when set, that the current packets are duplicated (default =0).

5. *Optimize (ON)* flag: it is in the Route caches' mappings (default=3).

   If a node has, in its Route cache, a mapping for a mobile node MN with ON equal to:

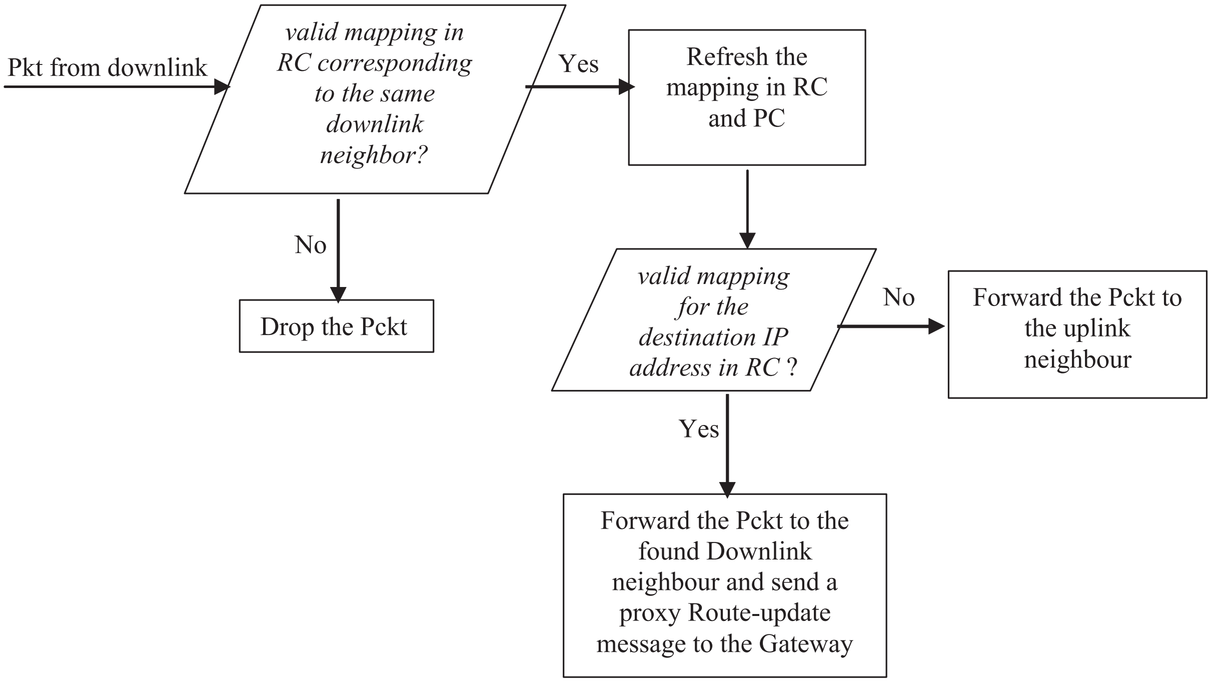

Pkt from downlink
valid mapping in RC corresponding to the same downlink neighbor?
Yes
Refresh the mapping in RC and PC
No
Drop the Pckt
valid mapping for the destination IP address in RC ?
No
Forward the Pckt to the uplink neighbour
Yes
Forward the Pckt to the found Downlink neighbour and send a proxy Route-update message to the Gateway

Figure 6.3: Optimized Uplink Routing Mechanism in Cellular IPv6 Network (RC and PC are the Route and Paging caches)

- 0, then the node "prepares" itself to be an optimizing node for the mobile node and its corresponding node. The ON is set to 0 during the handover establishment. The node in question does not send data through the optimal path as long as ON is equal to 0, in order to prevent the packets routing to the new cell during handover.
- 1, then the node can be an optimizing node for the MN and its corresponding mobile node. The optimizing node must send data through the optimal path.
- 2, then the data packets sent by the mobile node MN must be duplicated by the node if it is an optimizing node.

6.4 Proposition and Enhancements

The Cellular IPv6 protocol uses two parallel cache systems in order to have faster lookup and better scalability. In fact, since only a portion of the mobile nodes will be in active state at any given time, it is better to separate the caches for active and idle mobile nodes. Following the same reasoning, we propose to apply two parallel uplink routing mechanisms, namely the non-optimized uplink routing mechanism specified in [SGCW00] and the optimized uplink routing mechanism that we detail in subsection 6.5.

When the MN wants to send packets, it checks the subnet prefix of the CN's address.

- If the IPv6 subnet prefix of the MN's CoA is equal to the IPv6 subnet prefix of the CN's CoA, then the MN sets the flag OR, as defined in section 6.3, to 1. As a result, the CIP nodes will apply the optimized uplink routing mechanism.

- Otherwise, the MN sends the packets without setting the flag OR. Then, the CIP nodes will route the packets with the non-optimized uplink routing mechanism.

This routing mechanism distinction will prevent the CIP nodes from scanning the caches mappings at each packet reception. Consequently, this will decrease the processing load in the nodes.

The optimized uplink routing proposed in [SMG$^+$01] was made at the expense of an increase of the signaling load. In fact, in order to retain the routing cache consistency, the optimizing node must send a “proxy route-update” packet at a rate just faster than the route-timeout. In our mechanism, we propose to reduce this signaling load.

The loss of packets during the transition between networks should be minimal. It is shown in some research studies that buffering packets improves the global performance of Mobile IP. Our study defines a buffering mechanism that attempts to meet this goal for Cellular IPv6.

6.5 Optimized Uplink Routing mechanism

When a data packet arrives to a Cellular IPv6 node, the following events occur (figure 6.3):

If the data packet arrives from a downlink neighbor, the Route cache entry of the source IP address is searched first in the current node's Route cache. If the data packet arrives from a different neighbor than the one that is in its mapping or no mapping exists for the IP address, then the packet is dropped.

Otherwise, if the data packet is coming from the same neighbor, the mapping is refreshed in the Route and Paging caches.

The current node checks if the destination IP address has a valid mapping in the Route cache. There are two cases:

1. If such a mapping exists, the packet is forwarded to the downlink neighbor found in the mapping. In this case, the current node becomes an optimizing node for the two communicating mobile nodes.

2. Otherwise, if the Route Cache contains no mapping for the destination IP address then the packet is forwarded to the uplink neighbor.

In the first case, the optimizing node generates a proxy route-update message towards the gateway. The proxy route-update contains the IP address of the MN in the control field of the Hop-by-Hop Options extension header. The sending rate of the proxy route-update in [SMG$^+$01] is controlled by a time interval shorter than the route-timeout interval (9s). Thus, more signaling load is generated than in the non-optimized uplink routing. We propose that the optimizing node sends the proxy route-update message once: this message is sent when the node receives the first data packet that causes the node to become an optimizing node.

The proxy route-update sets the ON, in the mapping corresponding to the MN, to 1. If it finds ON equal to 2, no change in the mapping will be done.

Moreover, the proxy route-update will set the expiration time in the MN's mapping to the sum of the current time and an estimated time. The latter will be estimated according to the dwell time of the mobile in the CIP domain (according to the speed of the mobile) and to the call duration (according to the traffic type). In this case, the CIP nodes will operate in a hard state. This will help to reduce the signaling load and will not affect the route cache consistency.

With our mechanism, the caches will always be updated according to the mobiles' movement, and the packets will arrive to the correct destination. In fact, a change in the route cache occurs:

- When one of the two communicating mobiles leaves the CIP domain. If the sending mobile node leaves the cell, then it sends a "paging-tear down"[1]. The latter removes the mappings corresponding to the sending mobile node in the caches. Whenever the receiver leaves the cell, then it also sends a paging-tear down. The latter removes the mappings corresponding to the receiver in the caches. The packets sent to the receiver, in this case, will pass through the gateway in the uplink direction. And these packets will refresh the caches.

- When the sending mobile node's traffic is forwarded uplink. In this case, the packets will refresh the caches of the optimizing node and will be sent to a new optimizing node.

- At the end of the session and after the expiration of an "idle time", the sending mobile node becomes idle. The mapping of the sending mobile node stored in the Route caches will expire. As a result, the optimizing node will be a regular Cellular IPv6 node. The sending mobile node sends frequent paging-update packets. These packets will refresh the Paging caches of all the nodes in the path leading to the gateway.

As a result, the caches are refreshed according to the mobility of the flows, to the session duration, ... Therefore, the proxy route-update can be sent less frequently without affecting the routing mechanism.

In order to retain the routing cache consistency, the optimizing node of a communicating nodes pair, when receiving the update messages coming from the sending mobile node, must refresh its caches. Nevertheless, it must prevent sending these messages upwards in order to prevent the refreshing of the caches belonging to the branch leading to the gateway. Otherwise, the expiration time in the caches' mappings will change from the estimated time value. In this case, this expiration time will be set to the sum of the current time and the Route-update time. This will cause the caches' entries to expire because the caches do not receive frequent proxy route-update messages.

6.6 Handover Handling

Not all wireless technologies have simultaneous connection capability, i.e. they cannot listen to the current BS while sending a route-update packet to the new BS. For this situation an indirect semi-soft handoff is used in Chapter 3. We propose to enhance the indirect semi-soft handover handling while taking into account the optimized uplink routing mechanism.

Our study defines a buffering mechanism for the indirect semi-soft handoff. This buffering mechanism reduces the loss and the packet delay during the handover. The key idea is that the optimizing node duplicates the packets destined to the moving receiver. The original packets will be routed via the optimal path. As for the duplicated packets, they are sent to the crossover node where they are stored. These duplicated packets will be routed to the new mobile location

[1] A paging-teardown packet is an IPv6 packet with a Hop-by-Hop Options extension header where the source address is the IP address of the sending mobile node, the destination address is the Gateway and the Hop-by-Hop option is of Paging-teardown type [SGCW00].

after handover. In this way, the delay and the packets loss will be optimized.

We assume that a call is set-up between a mobile node and a corresponding node. For the ease of the handover handling description, we will limit our protocol exchange from a sending mobile node to a receiving mobile node.

When a mobile node performs handover, the following sequence of events occurs:

1. The mobile node sends a route-update packet to the current BS. This packet has the IP address of the new BS as destination IP address. The route-update packet contains the address of the corresponding mobile in the control field of the Hop-by-Hop Options extension header. The I flag is set to indicate the indirect semi-soft handoff.

   We distinguish between two cases:

   If the mobile node is a sending mobile node, then it sets the OR flag of the route-update packet to 1.

   If the mobile node is a receiving, then it sets the OR flag of the route-update packet to 0. This implies that the packet will reach the gateway. In this way, the route-update packet will reach the crossover node. This will not happen when the optimized uplink routing mechanism is used (OR=1) and when the optimizing node is hierarchically under the crossover node.

   The current BS forwards the route-update packet to the Gateway. The latter uses then normal IP routing to deliver the packet to the new BS. The route-update packet sets the flag ON to 2 in the Route caches mappings that correspond to the sending mobile node. This is done to all the nodes belonging to the branch starting from the current base station up to the gateway.

   In the following steps, we consider that the mobile node that is performing handover is the receiver. This is an important case, since the packets sent to the receiver must be duplicated in order to prevent the packets loss. When the mobile node that is performing handover is the sending mobile node, there is no need to perform packets duplication.

2. When the new BS receives the indirect semi-soft handoff packet, a semi-soft route update packet is created (I=0, S=1) with the IP address of the mobile node as the source address. It is then forwarded upstream. The semi-soft route-update packet creates new mappings in the Route and Paging Cache similarly to regular route-update packets. However, it sets the flag ON in the Route cache mapping that corresponds to the sending mobile node to 0. This is done to prevent the routing of packets to the new cell before the handover takes place. Recall that we are proposing to study the indirect semi-soft handoff.

   When the semi-soft route-update packet reaches the crossover node where the old and new path meet, the new mapping is added to the cache instead of replacing the old one.

3. Packets that are sent to the mobile receiver must pass by the optimizing node. When the flag ON, in the mapping corresponding to the sending mobile node, is equal to 2, the optimizing node performs the duplication of the packets (figures 6.4). The original packets are routed using the optimizing routing mechanism. As for the duplicated packets, they are routed towards the crossover node and stored in the buffer located in the crossover node.

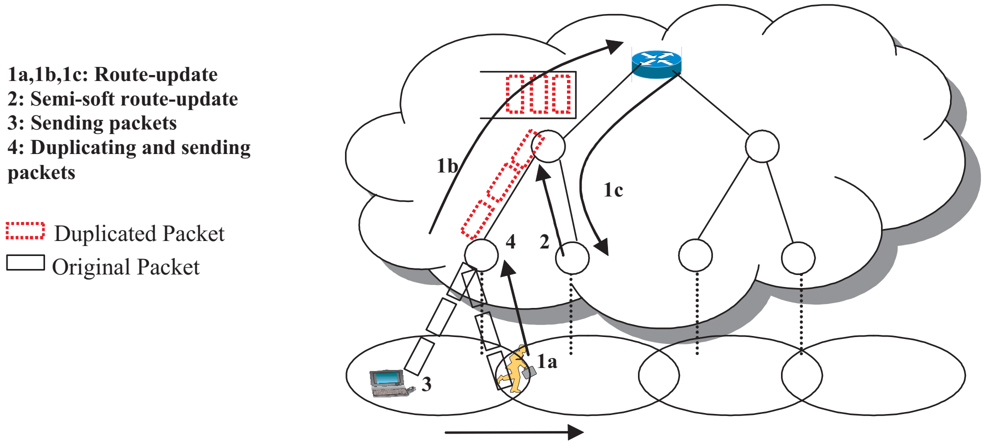


Figure 6.4: Packets Duplication During Handover

Consequently, even during handover, the packets are always sent via the optimal route, and the duplicated packets wait in the crossover node in order to be sent via the shortest path.

It is noteworthy that the duplicated packets do not refresh the caches in our proposition. Otherwise, the caches must be refreshed each route-update time: this would incur more signaling load. Thus, the flag DUP of the duplicated packets will be set to 1: the CIP nodes will be able to identify the duplicated packets.

4. When the mobile node moves to the new cell, it sends a route-update packet (OR=1, I=S=0) to the current BS. This packet has as a destination IP address, the IP address of the new BS. The packet in question contains the address of the moving mobile and the corresponding node's address in the control field of the Hop-by-Hop Options extension header. The route-update packet sets ON to 3 in the Route cache mapping corresponding to the sending mobile node, if it finds ON equal to 2. The current BS will then forward this packet to the old base station.

5. When receiving the route-update packet, the old base station sends a paging-teardown packet (OR=0, I=S=0) with the IP address of the receiver in the source address. This packet contains the source address of the mobile and the corresponding mobile address in the control field of the Hop-by-Hop Options extension header. This paging-teardown removes all the mappings concerning the moving mobile in the Caches except for the ones pointing to the new Base Station.

   When the paging-teardown arrives to the crossover node, it forces the buffer to free the packets sent to the receiver (figure 6.5). The freed packets take then the optimal path to arrive at the new mobile location.

6. The mobile node sends a remove-mapping message (OR=0, I=S=0) that contains the address of the corresponding mobile. This message will set ON in the mapping corresponding to the sending mobile node to 3, if it finds it equal to 2. This is done in all the caches of the CIP nodes belonging to the branch leading to the gateway.

   The handover is then complete.

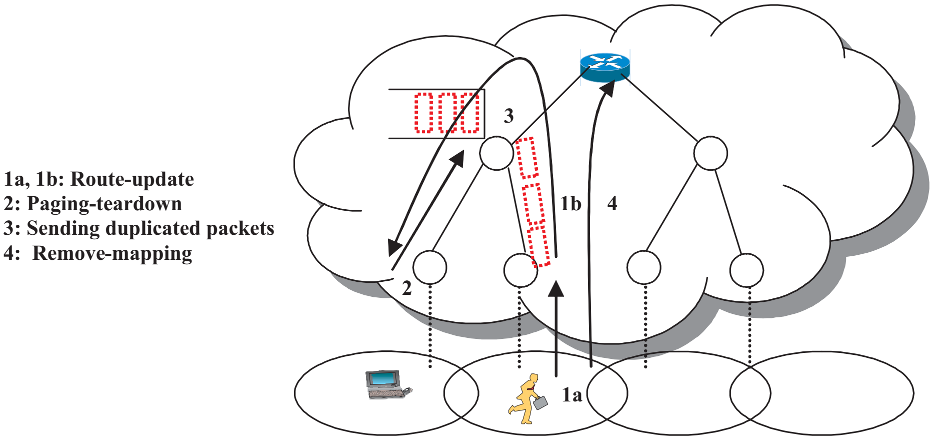


Figure 6.5: Packets Reception After Handover

6.7 Simulation

6.8 Numerical Results and Performance Analysis

In order to study the performance of the proposed mechanism, simulations were carried out using OMNeT++ [OMN]. The CIP network illustrated in figure 6.2 was simulated. The model parameters were chosen to study the proposed mechanism and to compare it to other mechanisms. We suppose that the wireless bandwidth is equal to $1Mb/s$, the wired link capacity dedicated for data is $1,92Mb/s$, that for the signaling is $128Kb/s$. The data traffic considered represents a typical WWW session (UDD $64Kb/s$) that consists of a sequence of packet calls. The parameters and laws that model the data traffic are specified in chapter 1. Mobile users are considered as pedestrians with mean speed of $1.8Km/h$ moving within the cells of radius $0,1Km$. Moreover, we consider that each CIP node has 3 buffers:

- the first buffer is dedicated for the signaling packets Bs.
- the second buffer is allocated to the data packets Bd.
- the third buffer is for the duplicated data packets during the handover Bh.

Since the signaling packets are important in Cellular IP networks, the signaling buffer Bs is allocated a percentage of the link capacity. Thus, Bs does not suffer from the resource contention. One better alternative is to apply the Round-Robin mechanism between the signaling buffer and the data packets. This would improve the bandwidth use but it would not change our results nor the conclusions of our study.

We apply the Head Of the Line (HOL) discipline with no-preemption in order to schedule the packets in Bd and Bh. The HOL serves the packets, stored in Bh first, after the paging-teardown reception. Note that the freed packets are the ones sent to the moving receiver which address is the source address of the paging-teardown.

Figure 6.6 shows the mean data load measured on the gateway interfaces. The mean load is the average of the load that we retrieve during an interval of time. One can see that the optimized routing mechanism lowers the load on the gateway which is considered as a bottleneck in the

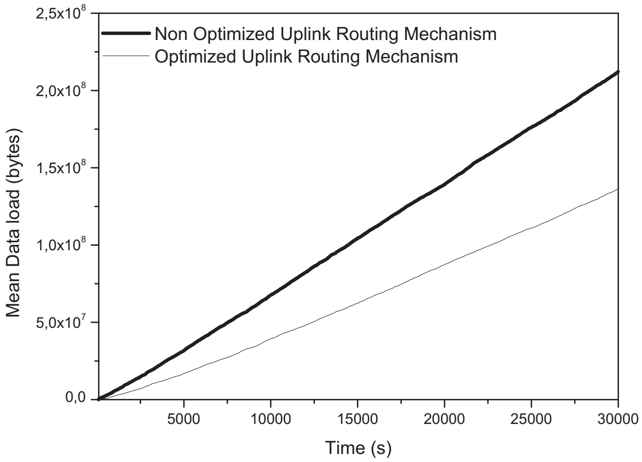


Figure 6.6: Mean Data Load

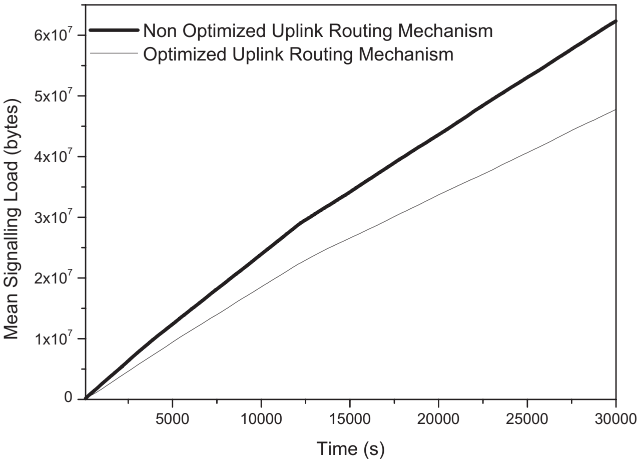


Figure 6.7: Mean Signaling Load

CIP network. Unlike [SMG$^+$01], the signaling load on the gateway interfaces is less than the one obtained with the non-optimized uplink routing mechanism (figure 6.7). In fact, the optimizing nodes retain all the update packets sent by the sending mobile node. This can alleviate the signaling load on the gateway. On the other hand, we argue that the proxy route-update packet must be sent once and that the expiration time, related to this packet, in the caches mappings must be well-chosen. Moreover, most of the signaling messages needed to establish the handover are sent through the optimal path. As a result, the decrease of the signaling load on the gateway interfaces is a logical consequence of our mechanism.

Figure 6.8 shows the mean number of hops crossed by the packets before arriving at destination. As we can see, this number of hops is constant with the non-optimized uplink routing mechanism. In fact, the packets always pass by the gateway before being routed to destination. Thus, the number of hops depend on the network topology. With the optimized uplink routing mechanism, the number of crossed hops is reduced. This is because the packets take the optimal path before arriving at destination.

As for the delay experienced by the packets, the curves depicted in figure 6.9 show better results than with the non-optimized uplink routing mechanism, due to the reduced number of hops crossed by the packets. This delay is also due to the duplicated packets that are received by the mobile upon sending the paging-teardown. As we can see, the localization of the storing buffer on the crossover node and not on the gateway helps to reduce the delay encountered by the packets.

Figure 6.10 depicts the delay of the packet sent during the handover. It can be seen that this delay is much higher with the non-optimized uplink routing mechanism. In fact, with our enhanced mechanism, the packets are sent through the optimal path during and after handover.

We also measured the mean establishment delay of handover in figure 6.11. We found that this delay is higher with the non-optimized uplink routing mechanism. In fact, with the optimized uplink routing mechanism and when the mobile moves to the neighboring cell, the path taken by

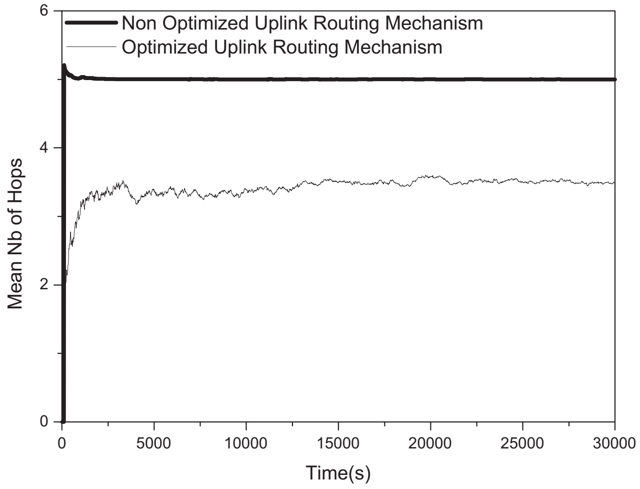


Figure 6.8: Mean Number of Hops

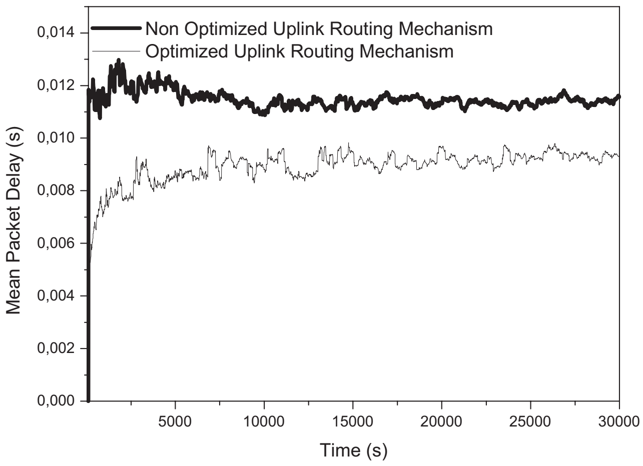


Figure 6.9: Mean Data Packet Delay

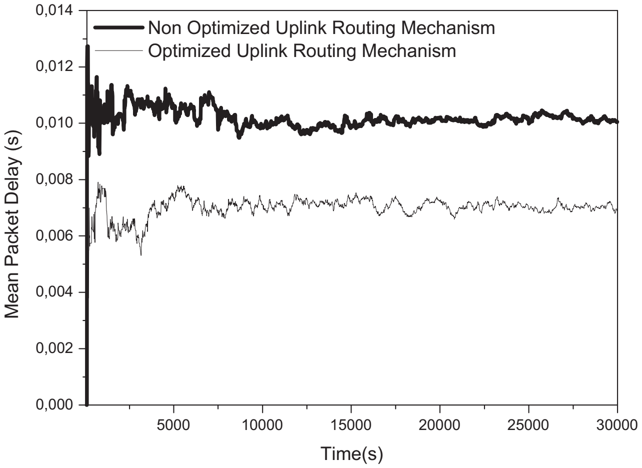


Figure 6.10: Mean Data Packet Delay During HO

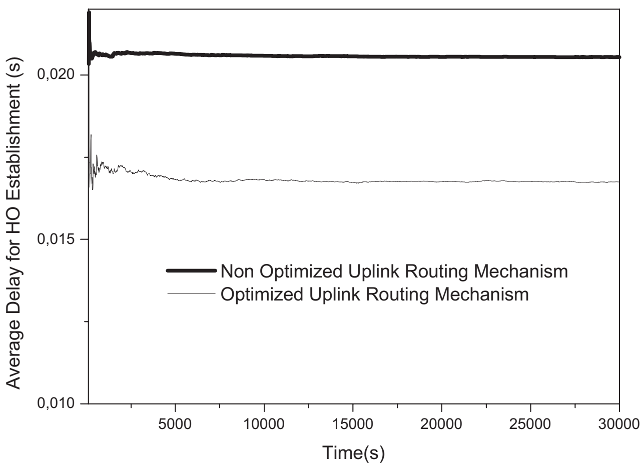


Figure 6.11: Mean HO Establishment Delay

the route-update packet is shorter than that with the non optimized uplink routing mechanism.

The retrieved simulation results show the benefits of the enhanced optimized uplink routing mechanism. These better results were obtained at the expense of some complexity added to the CIP nodes. One must make the trade-off between better performance and complexity.

6.9 Conclusion

This chapter presents an enhanced uplink routing mechanism coupled with a smooth handover study. This study is made to enhance the performance of the network in the case of intra-network traffic.

Our proposal aims to minimize the delay and the loss experienced by the data packets during communications and especially during handover. It also reduces the signaling load on the network gateway. It is noteworthy to indicate that the proposed uplink routing mechanism behaves better with a deep toplogy tree. Simulation results show the good performance obtained at the expense of some complexity added to the Cellular IPv6 nodes.

In an attempt to provide QoS for users generating inter-network traffic, we investigate, in the next chapter, the issue of providing an efficient CAC that offers QoS to different classes of service. Moreover, we propose an end-to-end QoS framework with Cellular IPv6 in the micro-mobility domains.

Chapter 7

End-to-End QoS Using Cellular IPv6

7.1 Introduction

Architectures for supporting QoS in the Internet can be broadly classified into two groups: the Integrated Services (IntServ) and the Differentiated Services (DiffServ). Both approaches suffer from drawbacks when used to provide resources for mobile nodes in wireless cellular environments. In fact, these approaches were designed in static environment and as a result, they are not fully adapted to the mobile environment, especially when Mobile IP is used as the mobility management protocol. Many works in the literature tried to enhance these approaches in order to provide QoS in mobile wireless networks.

Regardless of the approach adopted in the next wireless mobile network, we think that a suitable call admission control must be designed in order to handle the QoS requirements. We believe that the wireless medium requires a fundamental change in the expectations we have from the service and the level of quality of service provided by the network. Multimedia applications need to be adaptive, to renegotiate the service request and to deal with changing conditions. On the other hand, end systems must be network aware: they must be able to adapt the multimedia streams accordingly. This adaptive approach implies that the network and the application are responsible for providing QoS to the mobile users. Our dynamic adaptive architecture DYNAA provides this adaptive approach as detailed in chapter 5.

The present chapter is twofold. In a first step, we propose to apply the DYNAA architecture to the Cellular IPv6 network. Simulation and numerical results show the interest of our proposition.

In a second step, we propose an end-to-end QoS approach using IntServ in the Cellular IPv6 networks and DiffServ in the core network. The proposed end-to-end QoS approach is presented and discussed before concluding the chapter.

7.2 DYNAA and Cellular IPv6 Networks

7.2.1 General approach

Our approach consists of integrating the proposed DYNAA architecture into the Cellular IPv6 networks. Moreover, we want to compare three types of adaptive CAC that affect the mobility and the HO management.

1. The first type is *distributed*. It is locally performed in each cell. A distributed CAC alleviates the centralized load in the Cellular IPv6 gateway and is achieved in a distributed manner without any central control over multiple cells. A distributed radio resource monitoring mechanism and a local management of admission control are integrated into this scheme. This scheme may be used within ad hoc networks.

2. The second type is *centralized* in the Cellular IPv6 gateway. The gateway has a global view of the network and performs efficient load balancing at the expense of generating signaling traffic. This scheme is used within UMTS networks.

3. The third type of adaptive CAC is hybrid: it combines a *centralized* and a *distributed* CAC.

7.2.2 Design principles

Our approach consists of the following design principles:

1. Support for soft QoS. We think that the network should provide at least a minimum level of QoS to the users' applications. The applications should be able to adapt to the changing network conditions.

2. Minimum disruption of service. In wireless network admission, a decision is made on new and handover calls. Since forced call terminations due to handover failure have significant negative impacts on the user's perception of network reliability, handover requires special considerations. We strive towards a solution that would provide the minimum forced termination probability (P_{drop}) due to the user's movement.

3. Minimum changes to Cellular IPv6 protocol. We add minimum modifications to the defined protocol.

7.3 Extensions to Cellular IPv6 protocol

In order to integrate DYNAA into the Cellular IPv6 protocol, we add an intelligent entity, called the "Distributed Controller" (DC), within each cell as shown in figure 7.1 and a new control field, "the Adapted Bandwidth field", to the regular IP packets.

7.3.1 Distributed Controller (DC)

The Cellular IPv6 gateway is a bottleneck for the domain since all the packets must pass through it before arriving at their destination. The gateway can be logically divided into three building blocks:

- a regular Cellular IPv6 node.

- a gateway packet filter. This gateway filter reads the destination IP address. If this address is the gateway's address, the packet is forwarded to the gateway controller. If the packet carries control information, it is processed by the gateway controller. Whenever the destination address is not the gateway's address, the packet is forwarded to the Internet.

- a gateway controller that processes the control packets sent to the gateway.

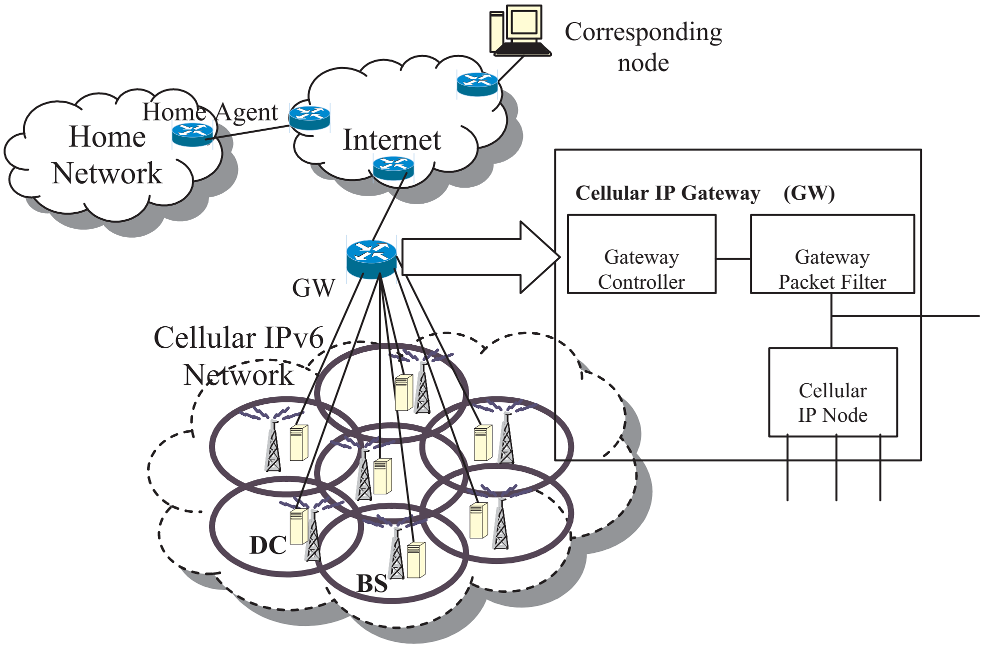


Figure 7.1: Cellular IPv6 network with the Distributed Controllers (DC)

In order to alleviate the load centralized in the gateway and to reduce the signaling traffic, we propose to decentralize the call admission control. Thus, we introduce a local entity called the Distributed Controller (DC) attached to each base station. The DC is responsible for the distributed call admission control. In the case of a centralized call admission control, the DCs distributed in the cells interact with the gateway in order to accept or refuse an incoming request.

7.3.2 Adapted Bandwidth field (AB)

When a call is accepted, other ongoing calls' bandwidth may be adapted in order to satisfy the incoming call. Thus, the mobiles having ongoing calls must be notified about the new granted bandwidth. This notification is done if the network decides that these ongoing calls have to adapt their bandwidth. In order to do that, the base stations must send packets containing the control information to the mobiles in question. This control information could be as well piggy-backed in data packets sent to the mobiles in question. Thus, we add a field in the Destination Options Header, namely the Adapted Bandwidth field (AB) such that:

$$AB = b_{i,j}$$

where i denotes the class of service, and $b_{i,j}$ is the bandwidth of an application having j layers and handled by the class i of service. This bandwidth must be allocated to the call that was set-up by the application in question. In this way, AB is the control information sent to the mobiles in order to notify them about the new granted bandwidth.

7.4 Centralized And Distributed Call Admission Control

We want to compare the performance of three types of CAC that rely on the DYNAA architecture, namely *the distributed*, *the centralized* and *the centralized/distributed* CAC schemes. It is noteworthy that the DYNAA architecture with its waiting scenario (DYNAA_Wait) is locally used in each cell for the three CAC schemes. Thus, the HO requests wait in queues in the HO areas, when the adaptation algorithms do not find enough resources for the HO requests. We consider two classes of service: the hard adaptive (HA) and the soft adaptive (SA) (chapter 5).

7.4.1 Centralized/distributed call admission control

This scheme combines a centralized and a distributed CAC. When a mobile requests admission to a Cellular IPv6 network, the DC attached to the base station decides whether to accept or not the incoming request. Thus the DC applies the adaptive CAC, defined within DYNAA (chapter 5).

Whenever the incoming request is a handover request and when the neighboring cell runs out of resources and cannot adapt other calls in order to satisfy the handover call, the handover request is inserted into a waiting queue. This is done in a transparent way; the gateway is not aware of the CAC performed within the DC.

The DC may adapt and even degrade other calls' bandwidth in the neighboring cell in order to satisfy the HO request. When the HO request is accepted, the base station sends a packet to the incoming mobile and to the mobiles that should adapt their calls' bandwidth in order to accept the HO request. This packet carries in the Destination Option Header the AB field. The AB field contains the new granted bandwidth to the mobile that the packet has been sent to.

Whenever the incoming request is a new call and if the DC refuses it, the base station sends a regular IP packet informing the mobile of the decision. This IP packet carries in the Destination Options Header the AB field. This AB field contains zero in order to notify the mobile of the refusal. When the new call request is accepted locally, the base station sends the request to the gateway that will apply the centralized CAC as explained in the next paragraph.

The gateway controller must have a global view of the domain load. Thus, it must retrieve some measurement reports from each cell. The gateway accepts more or fewer new calls according to the retrieved reports. The HO is controlled directly by the reservation made by the gateway controller.

Let us consider M cells belonging to a CIP network controlled by a gateway. Each cell is assumed to have C channels providing service to HA and SA users. We introduce a threshold-type CAC algorithm where C_1, C_2 are two thresholds such that $C_1 \leq C_2$.

When the DC of the serving cell accepts a new call, it sends the request to the gateway. The new SA (respectively HA) call is blocked by the gateway controller if the total number of ongoing calls in the domain exceeds the resource allocation C_1 (respectively C_2). In this way, we reserve guard channels for the HO requests distributed in the Cellular IPv6 domain. Note that the handover requests are not processed by the gateway. Instead, they are locally accepted or refused by the DC of the neighboring cell.

One interesting policy that could be used by the gateway is to allocate a channel from a non loaded cell, and thus performing a load balancing. This will be the subject of further study.

7.4.2 Centralized call admission control

This scheme behaves as the above proposed scheme. However, it is different at the local level. In fact, the DC attached to each base station applies the DYNAA architecture with the waiting scenario. However,the centralized call admission control does not adjust the calls' bandwidth to the HO load.

When there is a HO request, we always accept the call without checking whether the HO load is high or not. Moreover, we do not check whether the degradation parameters are less or higher than the upper-bound values. Thus, this scheme is not dynamic and the mobility handling is managed by the gateway through the threshold-type CAC algorithm.

7.4.3 Distributed call admission control

This proposed scheme is performed by each DC in a distributed manner: no central coordination is performed. The distributed admission control scheme relies on the adaptive CAC distributed in the base stations. Although, dynamically adapted to the flows mobility, this approach does not provide a global view of the network as in the centralized/distributed CAC.

7.5 Simulation Results

In order to compare the different CAC schemes, simulations were carried out. The network depicted in figure 7.1 was simulated. The cells under consideration are assumed to have a radius of $100m$ and a capacity C of 60 channels. The CIP domain thresholds C_1 and C_2 are equal to $336(80\%C)$ and $378(90\%C)$ respectively. Users are assumed to be uniformly distributed moving with an average speed of $4Km/h$. The unencumbered session duration of a HA call is exponentially distributed with mean $120s$. The arrival of a HA call in the simulation is assumed to be Poisson with mean rate λ_{nv}.

On the other hand, the SA class is represented by a typical WWW session (UDD $64Kb/s$) that consists of a sequence of packet calls (chapter 1). The HTTP session arrival process is Poisson with mean rate λ_{nd} such that $(\lambda_{nd} = \lambda_{nv} = \lambda_{ntot}/2)$, where λ_{ntot} is the total call originating rate density. The bandwidth requirements are $(b_{1,1}, b_{1,2}, b_{1,3}){=}(1, 2, 3)$ for class 1 (HA) and $(b_{2,1}, b_{2,2}, b_{2,3}){=}(2, 4, 6)$ for class 2 (SA) such that $(b_{1,req}, b_{2,req})$ $=(2, 4)$. Note that the upper-bound values of the degradation parameters DD, DR are respectively $(DR_{1,qos}, DD_{1,qos})$ $= (0.01, 0.01)$ for HA class and $(DR_{2,qos}, DD_{2,qos}) = (0.1, 0.1)$ for SA class. On the other hand the threshold values of P_{drop}, $(P_{drop\_min}, P_{drop\_max})$ are equal to $(0.0075, 0.009)$, and the required P_{drop} is $P_{drop,qos} = 0.01$. The Lth which is the parameter related to the QLT scheduling policy (see chapter 4) is equal to 2. The monitoring interval ΔT is taken to be equal to $5s$.

Figure 7.2 illustrates the new call blocking probabilities obtained with the three schemes. The new SA blocking probability B_{nSA} retrieved with the centralized and the centralized/distributed CAC is much higher than that with the distributed CAC scheme. In fact, the C_1 threshold blocks new SA calls in order to prioritize new HA calls and HO requests: this leads to a higher B_{nSA} and a lower new HA blocking probability B_{nHA}. However, with the high increase of the rate density, the network becomes so loaded that B_{nHA} significantly increases with the centralized and the centralized/distributed CAC and becomes equal to that of the distributed scheme. The

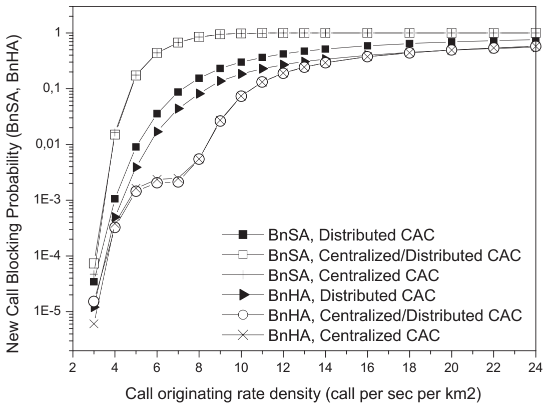


Figure 7.2: New Call Blocking Probability

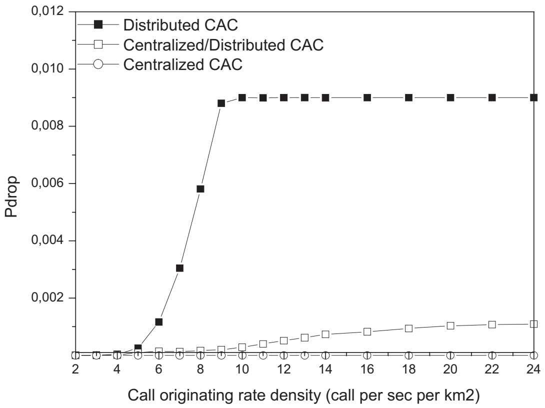


Figure 7.3: Forced Termination Probability

gateway becomes a real bottleneck with the load increase, in the centralized and the centralized/distributed scheme.

On the other hand, the new HA blocking probability B_{nHA} with the distributed CAC is higher than with the centralized and centralized/distributed CAC. In fact, with the centralized and centralized/distributed CAC, blocking SA calls at the C_1 threshold gives more chance to the new HA calls to be served until the C_2 threshold.

The forced termination probability is depicted in figure 7.3. It can be seen that the distributed scheme keeps the P_{drop} under the QoS requirement $(0,01)$. In fact, with DYNAA, we are concerned about keeping this probability less than $P_{drop,qos}$. On the other hand, P_{drop} with the centralized CAC is almost negligible: this is because the CAC deployed in the base stations accepts the HO requests even if the degradation parameters are higher than the upper-bound parameters.

However, P_{drop} with the centralized/distributed CAC scheme increases with the call originating rate density rate increase until the rate density attains the value of $19calls/s/Km^2$, after which it becomes stable at the value of 0.001. It can be seen that P_{drop} with the centralized/distributed scheme is less than that with the distributed CAC due to the C_2 threshold. In fact, the C_2 threshold prioritizes the HO requests by blocking the new HA and new SA calls whenever their number exceeds the C_1 and C_2 thresholds. This considerably reduces the forced termination call probability.

Figures 7.4 and 7.6 depict the HA degradation parameters retrieved with the three schemes. DD_1 and DR_1 obtained with the distributed CAC are less than those obtained with the centralized CAC. In fact, with the centralized CAC, more HA handover calls are accepted. Furthermore, with the centralized scheme the calls' bandwidths are always adapted in order to satisfy an incoming HO request. Thus, the probability of degrading the HA calls in the centralized scheme is greater than in the distributed scheme. This result is very frustrating when using the centralized scheme, since HA class is demanding in terms of degradation parameter.

With the centralized CAC, the C_1 threshold blocks new SA calls. Thus there are fewer new SA calls and fewer handover SA calls in the cell with the centralized scheme than with the distributed scheme. The SA degradation parameters (DD_2 and DR_2) are very small with the centralized scheme and tend towards zero (figures 7.5, 7.7).

With the centralized/distributed CAC, the very small dropping probability leads to the execution of the BNDA (Bandwidth with no Degradation Algorithm) algorithm almost all the time (chapter 5).Recall that this algorithm adapts the existing calls' bandwidth but does not degrade the calls when there are not enough resources for an incoming HO request. Hence, the degradation parameters (DD_1, DD_2, DR_1, DR_2) with the centralized/distributed CAC tend towards zero.

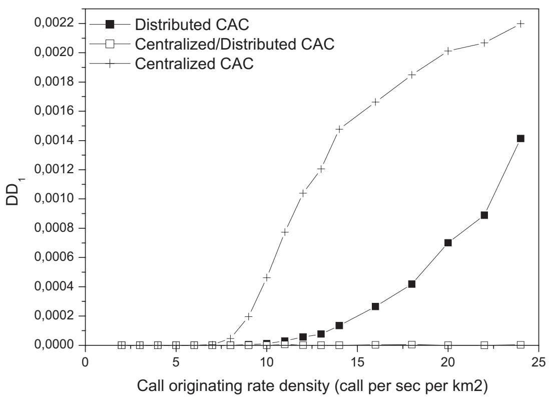


Figure 7.4: Mean HA Degradation Degree

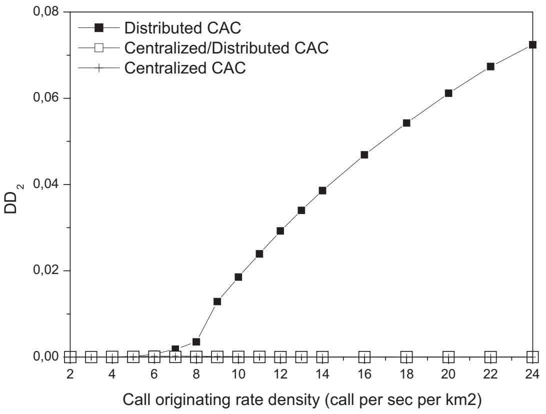


Figure 7.5: Mean SA Degradation Degree

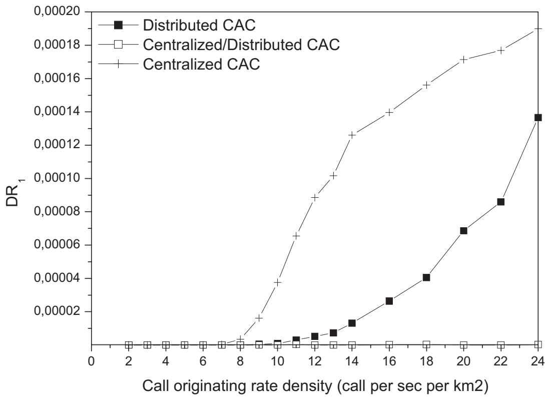


Figure 7.6: Mean HA Degradation Ratio

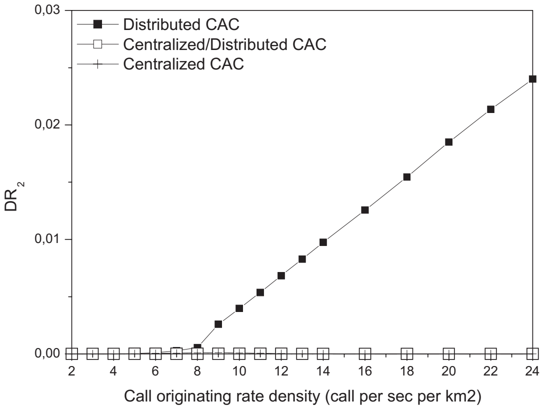


Figure 7.7: Mean SA Degradation Ratio

In conclusion, one can see that with the centralized/distributed CAC scheme, a small forced termination probability is achieved and a good new HA blocking probability is attained at the expense of a higher SA blocking probability. On the other hand, small degradation parameters are obtained. This leads us to conclude that the centralized/distributed scheme outperforms the other schemes, when comparing the QoS performance. Nevertheless, the performance improvement is achieved at the expense of loading the gateway by the requests and the signaling process.

On the other hand, the distributed scheme has good performance and reduces the load and the processing in the gateway. The distributed scheme can be efficient and flexible in the case of highly bursty traffic. Thus, there is a trade-off between the centralized/distributed and the distributed scheme.

Some open issues still remain to study. An end-to-end QoS support is an important issue to consider. The QoS architectures that are proposed for the Internet (DiffServ, IntServ) are not adapted to the mobile environments, especially when Mobile IP is used as the mobility management protocol. Next, the QoS support of IntServ and DiffServ in mobile environments is discussed, together with a proposition of an end-to-end QoS framework.

7.6 End-to-End Quality of Service

In the previous sections, we studied the QoS in the access network. Now, we discuss an approach for providing an end-to-end QoS, using Cellular IPv6 in the micro-mobility domains.

The main drawback of the current Internet is the lack of QoS support. QoS support is essential for real-time applications such as Internet Telephony and on-line video retrieval. During the last years, the Internet community spent many efforts to develop an Internet QoS architecture based on the Integrated Services architecture and RSVP. IntServ performs fine-grained and per flow QoS. However, the IETF RSVP working group stated that RSVP and the IntServ approach can not be deployed in large-scale Internet backbones due to scaling and billing problems.

Differentiated Services is a new approach for QoS support in the Internet. DiffServ provides coarse-grained and per-aggregate QoS. DiffServ uses a differentiation model where packets are classified into a relatively small number of classes at the network edge. The packets of each class are marked and traffic conditioned by the edge router, according to the resource commitment negotiated in the Service Level Agreement (SLA). The resource allocation is performed by the “bandwidth broker” in a centralized manner without dynamic resource reservation signaling. DiffServ involves pre-configuration of resources (for a specific class) along a path and the use of admission control algorithms at the edge to limit the offered load along the path. The rapid movement of an MN however leads to changes in the traffic path and makes it hard to devise effective admission control strategies under a static resource configuration regime.

Several problems occur when DiffServ is used in conjunction with Mobile IP. These problems can be classified into the following five categories: network provisioning in mobile environments, lack of dynamic configuration, definition and selection of service level agreement, mobile flow identification and billing.

Both DiffServ and IntServ approaches have been designed in the context of a static environment (fixed hosts and networks) and as a result, they are not fully adapted to mobile environments, especially when Mobile IP is used as the mobility management protocol.

In the subsequent sections, we outline the problems that exist with DiffServ, IntServ and RSVP along with the solutions proposed in the literature. Afterwards, we give our approach for an end-to-end QoS.

7.6.1 DiffServ and Mobile IP

An IP tunnel encapsulates IP traffic in another IP header as the traffic passes through the tunnel. The presence of these two IP headers is a defining characteristic of IP tunnels, although there may be additional headers inserted between the two IP headers. The inner IP header is that of the original traffic; an outer IP header is attached and detached at tunnel endpoints.

A primary concern for differentiated services is the use of the Differentiated Services Code Point (DSCP) in the IP header [NBBB98]. The DiffServ architecture permits intermediate nodes to examine and change the value of DSCP, which may result in the DSCP value in the outer header being modified between tunnel endpoints. When leaving the IP tunnel, packets will be decapsulated and the DSCP value in the outer header will be lost. Thus, there is a problem when an IP tunnel pass through a DiffServ domain.

An interesting model, called *the uniform model*, is presented in [Bla00]. This model views IP tunnels as artifacts of the end to end path from a traffic conditioning standpoint; tunnels may be necessary mechanisms to get traffic to its destination(s), but have no significant impact on traffic conditioning. *Implementations of this model copy the DSCP value to the outer IP header at encapsulation and copy the outer header's DSCP value to the inner IP header at decapsulation.* The use of this model allows IP tunnels to be configured without regard to the DiffServ domain boundaries because the DiffServ traffic conditioning functionality is not impacted by the presence of the IP tunnels.

7.6.2 IntServ and RSVP

The IntServ approach uses RSVP to explicitly signal and dynamically allocate resources at each intermediate node in the traffic path [ZDE$^+$02], [Wro97b], [BZB$^+$97]. Two types of messages, Path and Resv are used to setup resource reservation states on the nodes along the path between a sender and a recipient.

Initially, the sender sends a Path message to the recipient to find the path all the way from the sender to the recipient for a specific flow. The Path message is updated by each router encountered along the path. The collected information is then carried to the recipient.
Each sender periodically sends a Path message for each data flow it originates.

Every RSVP message carries a SESSION object which identifies a flow. The SESSION object contains the destination IP address of the flow, the protocol ID and the destination port number.

A Path message carries the following objects:

- *SENDER_ TEMPLATE:* it identifies the sender and consists of the sender IP address and optionally the UDP/TCP sender port number.
- *SENDER_ TSPEC:* it describes the traffic characteristics of the flow generated by the sender.

- *ADSPEC:* it describes the aggregate QoS characteristics of the path.
- *PHOP:* it identifies the previous hop which sent this Path message.

Each RSVP-capable node along the path traveled by the Path message captures this message and processes it to create path state for the sender defined by the SENDER_TEMPLATE and SESSION objects.

The recipient is in charge of taking a global decision on how much bandwidth to reserve, according to the information collected by the Path message. This information is then stored in a Resv message, and is notified back to each router along the path. This procedure is an end-to-end one. Upon receiving the Resv message, each router on the path will reserve resources for the specific flow if sufficient resources are available.

A Resv message contains a FLOWSPEC object. A FLOWSPEC object consists of two sets of numeric parameters:

- *RSPEC:* it defines the desired QoS
- *TSPEC:* it describes the traffic characteristics of the data flow.

RSVP takes a “soft state” approach to manage the reservation state in routers and hosts. The RSVP soft state is created and periodically refreshed by Path and Resv messages. The sate is deleted if no matching refresh arrives before the expiration of a “timeout interval”. The state may also be deleted by an explicit teardown message.

RSVP teardown messages remove the path or reservation state immediately. There are two types of RSVP teardown messages:

1. PathTear messages are initiated explicitly by senders or by path state timeout in any node, and they travel towards all receivers. PathTear messages delete path state as well as dependent reservation state along the way.
2. ResvTear message are initiated explicitly by the receivers or by any node in which reservation state has timeout and they travel towards all matching senders. They delete reservation state along the way. A PathTear or ResvTear message may be conceptualized as a reversed-sense Path or Resv message respectively.

Besides the above-mentioned messages, there are three different RSVP messages:

- PathErr messages report errors in processing Path. They travel towards senders and are routed hop-by-hop using the path state.
- ResvErr messages report errors in processing Resv or they may report the spontaneous disruption of a reservation. They travel towards receivers and are routed hop-by-hop using the reservation state.
- ResvConf messages are sent to acknowledge reservation requests. To request confirmation for its reservation request, a receiver includes in its Resv message a confirmation-request object containing the receiver’s IP address.

7.6.3 IntServ-RSVP drawbacks

The IntServ-RSVP is mainly criticized because of the cost of state maintenance and processing overhead in each router. Moreover, all the routers must support the RSVP signaling protocol. As a result, the IntServ limits the scalability.

On the other hand, RSVP cannot be used directly in a mobile environment for two reasons:

1. RSVP messages are "invisible" to the intermediate routers of an IP tunnel in Mobile IP because the IP tunnel is implemented by an IP-in-IP encapsulation scheme. In fact, Path and Resv will be encapsulated in an IP-in-IP encapsulated packets with 4 a protocol number in the outer IP header, concealing the original RSVP protocol number 46 in the inner IP header. As a consequence, the routers on the path of an IP tunnel cannot correctly recognize RSVP messages to provide the required QoS.

2. RSVP is not aware of mobility. According to the original signaling protocol, the resource reservation cannot be dynamically adapted along with the movement of a mobile node. In other words, once a mobile node hands off to a new region, its prior reserved resources are no longer available and the service quality of the mobile may degrade significantly due to the lack of resources reserved for the mobile in the new region.

In an attempt to solve the problem of the RSVP integration into a mobile environment, many works tried to enhance the RSVP in order to provide QoS in mobile wireless networks. These works are briefed in the next subsection.

7.6.4 RSVP in mobile environments

Several studies were proposed to deal with the incompatibility of RSVP with the mobile environments.

Terzis et al. in [TSZ99] propose the RSVP Tunnel to resolve the RSVP message invisibility problem. The underlying principle of RSVP tunnel is to establish nested RSVP sessions between the tunnel end-points, namely entry and exit points. That is, *an extra pair of tunnel Path and Resv messages*, without encapsulating IP headers, is sent to establish a QoS guaranteed communication path between the tunnel entry and exit points. The RSVP Tunnel proposition can solve the RSVP invisibility problem. However, with this proposition there is no solution for reducing service disruption due to frequent mobility of a mobile node.

Talukdar et al. propose MRSVP [TBA01], a solution in which reservations are pre-established in the neighboring access routers and thus, the mobile does not need to re-establish them. To achieve this, proxy agents are introduced and a distinction is made between *active* and *passive* reservations. A reservation for a flow on a link is called *active*, if data packets of that flow are traveling over the link to a receiver; a reservation is *passive*, if resources are reserved for the flow on the link, but actual data packets of the flow are not being transmitted over the link. Although this proposal solves the timing delay for QoS re-establishment with other reservation, several disagreements are introduced. A drawback of MRSVP is that it relies on the mobile node to supply its mobility specification (i.e a list of care of addresses in the foreign subnets it may visit). On the other hand, this scheme underlies an over-reserving resources of a scarce medium such as wireless bandwidth.

In an attempt to enhance the excessive resource reservations that waste bandwidth and degrade the network performance, Tseng et al. propose in [TLL01] a "hierarchical MRSVP protocol". Using this protocol, resources are reserved only when a mobile node resides in the overlapping area of boundary cells of two regions. This proposal outperforms MRSVP in terms of reservation blocking, forced termination and session completion probabilities.

Chen et al. [CH00] propose an extension of RSVP based on IP multicast. RSVP messages and actual IP datagrams are delivered to a mobile node using IP multicast routing. With this approach, the mobile proxy is an entity that acts as a proxy agent for a mobile node to make various reservations along the data path, in the current cell and the surrounding cells. The mobility of a node is modeled as a transition in a multicast group membership. The multicast tree is modified dynamically every time a mobile node is roaming to a neighboring cell. A new mobile node joining a multicast group results in establishing a new branch in the multicast tree. Once these new branches are formed, path messages from the sender are forwarded to mobile proxies along the multicast tree. Upon receiving the path messages, a "conventional reservation" message from current mobile proxy and " predictive reservation" messages from neighbor proxies are propagated towards the sender along the multicast tree.

In this approach, service degradation and packet delay are minimized and re-routing of flows is eliminated. However, this approach generates overload to manage multicast tree dynamically and network resources are poorly managed.

Chiruvolu et al. in [CAV99] proposed a Mobile IPv6 and RSVP integration model. The main idea of Mobile IPv6 and RSVP inter-working is to use RSVP to reserve resources along the direct path between the CN and the MN without going through their HAs since Mobile IPv6 supports route optimization. In this model, resources are initially reserved between the CN and the MN's original location. Whenever the MN performs a handover, a new RSVP signaling process is invoked immediately to reserve resources along the new path. For each handover, the MN as receiver has to wait for a new Path message from the CN; only after getting the Path message, it can issue a new Resv message to the CN. All these RSVP negotiations are conducted end to end even though the path change may only affect a few routers within the whole path during a single handover. Thus, the long handover resource reservation delays and large signaling overheads caused by this end-to-end RSVP renegotiation process could lead to notable service degradation in providing real-time services.

7.7 An End-to-End QoS Proposal

7.7.1 Co-existence of IntServ and DiffServ

The IntServ-RSVP has been criticized because of the cost of state maintenance and of the processing overhead in each router that limit scalability. However, in a micro-mobility domain where the number of nodes is limited, RSVP can be used without causing problems to scalability.

As for DiffServ, it is much more scalable in an architecture that supports low mobility. Thus, we propose to apply the IntServ-RSVP approach in the micro-mobility domains and DiffServ in the core network (figure 7.8).

The DiffServ domain behaves as a "non-RSVP cloud" towards RSVP [BZB$^+$97]. Path and Resv messages do not guarantee any QoS level in DiffServ domains.

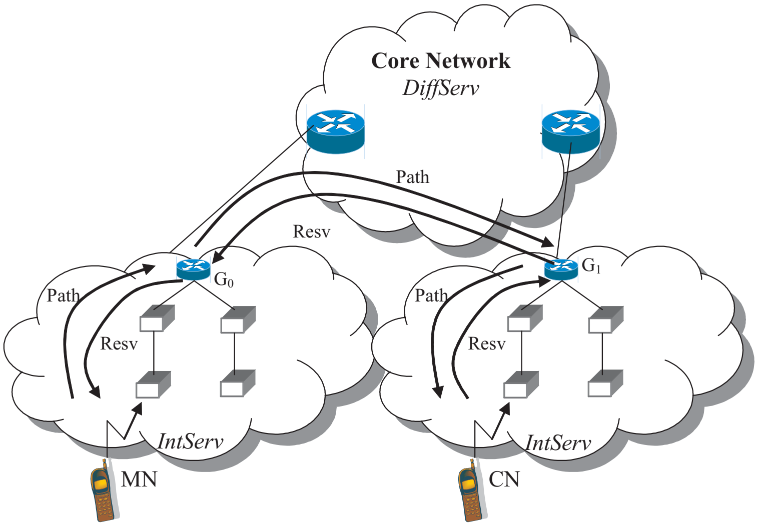


Figure 7.8: Co-existense of DiffServ and IntServ

Both RSVP and non-RSVP routers forward Path messages towards the destination address using their routing tables. Therefore, the routing of Path messages will be unaffected by non-RSVP routers in the path. When a Path message crosses a non-RSVP cloud, it carries to the next RSVP-capable node the IP address of the last RSVP-capable router before entering the non-RSVP cloud. A Resv message is then forwarded directly to the next RSVP-capable router on the path back towards the source

IntServ and DiffServ domain can coexist and offer an end-to-end QoS. The ingress[1] and egress[2] nodes in the DiffServ domain must establish the correspondence between a class of service in the IntServ domain and a class of service in the DiffServ domain.

Within the DYNAA architecture, we considered two classes of service: the hard adaptive and the soft adaptive. We propose the following mapping of the classes of service:

Class of Service in DYNAA	**IntServ Class**	**DiffServ Class**
Hard Adaptive	Guaranteed Service	Expedited Forwarding
Soft Adaptive	Controlled Load	Assured Forwarding 1 (AF1)

Table 7.1: Classes Mapping

We further assume that the RSVP signaling messages are mapped to the Expedited Forwarding class (EF).

Another proposition for the mapping can be as follows:

With this proposition, the RSVP signaling messages are mapped to the Assured Forwarding 1 (AF1). This proposition will make it easier to the DiffServ to handle the different classes of service.

[1] An ingress node is a DiffServ boundary node that handles traffic as it enters a DiffServ domain

[2] An egress node is a DiffServ boundary node that handles traffic as it leaves a DiffServ domain

Class of Service in DYNAA	**IntServ Class**	**DiffServ Class**
Hard Adaptive	Guaranteed Service	Assured Forwarding 2 (AF2)
Soft Adaptive	Controlled Load	Assured Forwarding 3 (AF3)

Table 7.2: Classes Mapping (bis)

7.7.2 Active and passive reservations

In our study, we adopt the concept of *active* and *passive* reservations as in [TBA01]. This concept is used to obtain high utilization and to guarantee successful mobility in the network. Note that the resources of passive reservations can be used by other flows requiring weaker service guarantees.

Unlike MRSVP, which establishes excessive passive reservations on all the MH's surrounding cells regardless which cell the mobile node is currently visiting, we consider that the passive reservations are made when the MN is in the overlapped area of two cells and when it is initiating a handover.

To achieve the active and passive reservations, we add two messages to RSVP [TBA01]. These are:

- Passive Path message: this message carries a SENDER_TSPEC for passive reservation
- Passive Resv message: this message carries a FLOWSPEC for passive reservation

In order to establish an end-to-end QoS, the three following messages are needed [TBA01]:

- Receiver_Spec: this message is sent by a mobile receiver. It carries the FLOWSPEC and the flow identification (i.e. the SESSION object) to its remote proxy agent.
- Receiver_MSpec (G,P): this message is sent by a mobile node to the appropriate node who sets up the routes of active and passive reservations. It contains the address of the gateway G of cellular IP domain that will be visited by the mobile node and the address of the remote proxy agent P of the mobile.
- Sender_Spec: a mobile sender uses this message to send its SENDER_TSPEC, ADSPEC and the destination address of a flow to a proxy agent.

7.7.3 Cell description

In each cell, we assume that there is a Distributed Controller attached to the base station as precised in section 7.3. The distributed controller is involved in the centralized/distributed CAC performed in the Cellular IPv6 network.

Besides the distributed controller, each cell contains a proxy agent. The proxy functionality is important for detecting and stopping certain messages in order to reduce signaling overload in the network. The proxy agent can make passive reservations on behalf of the mobile nodes which are not currently present in the proxy agent's cell.

The proxy agent is attached to the base station. This place limits RSVP signaling messages. The proxy agent at the current cell of a mobile node is called *local proxy agent*; the proxy agent at the next cell is called *remote proxy agent*. The remote proxy agent will make passive reservations on behalf of the mobile node. The local proxy agent of a mobile node acts as a normal router for the mobile node and an active reservation is set up from the sender to the receiver via its local proxy agent.

A proxy agent for a mobile receiver (sender) sends Resv (Path) messages periodically for reservations. To reduce the passive reservation overhead, the refresh interval for Passive Resv (Passive Path) messages can be made at least twice that of Resv (Path) messages sent for an active reservation.

The refresh messages do not reach the end nodes. *This avoids the overload problem on the air interface between the mobile nodes and the base station.*

Note that the mobile node can identify the proxy agent's IP address when listening to the beacon signal. We assume that the beacon signal contains the IP address of the proxy agent.

In the following subsections, we explain in detail the signaling messages needed to provide an end-to-end QoS during handover.

7.7.4 RSVP signaling during handover

When a MN moves between base stations in wireless access networks, RSVP signaling will only be needed between the crossover node, being aware of the former RSVP path and the remote proxy agent. Recall that the crossover node is at the intersection of two paths: one path is from the gateway to the previous base station and the second path is from the gateway router to the new base station.

The negotiation about resources can be managed efficiently in a small part of the network instead of traveling the entire path between the sender and the receiver.

We distinguish two cases:

1. The mobile node MN is the receiver
2. The mobile node MN is the sender

7.7.4.1 Mobile node is the receiver

Intra-domain handover: When the MN is moving within the overlapping area of the cells C0 and C1 belonging to the same Cellular IP domain, an intra-domain handover is initiated. The following events occur (figure 7.9 and figure 7.10):

1. The MN sends a route-update packet to the new base station. The MN also sends a Receiver_Spec to the remote proxy agent P1 in order to specify the desired QoS parameters.
2. When the new base station receives the route-update packet sent by the MN, the DC attached to the new base station performs the CAC and the bandwidth adaptation algorithms. If the HO request is accepted, the new base station generates a semi-soft route-update packet. This semi-soft route-update packet creates new mappings in the Route and Paging Caches similarly to regular route-update packets. When the semi-soft route-update packet reaches the crossover node where the old and new path meet, the new mapping is added to the cache instead of replacing the old one.
3. Upon receiving the semi-soft route-update message, the RSVP module in the cross node sends a Passive Path message to the remote proxy agent P1. The Path message can be

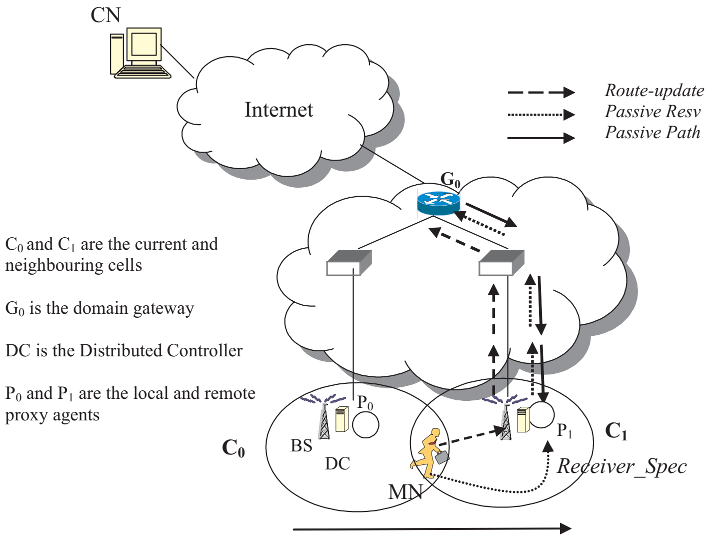

CN
Internet
Route-update
Passive Resv
Passive Path
G_0
C_0 and C_1 are the current and neighbouring cells
G_0 is the domain gateway
DC is the Distributed Controller
P_0 and P_1 are the local and remote proxy agents
P_0
BS
DC
C_0
MN
P_1
C_1
Receiver_Spec

Figure 7.9: Intra-domain Handover: the mobile node MN is a receiver

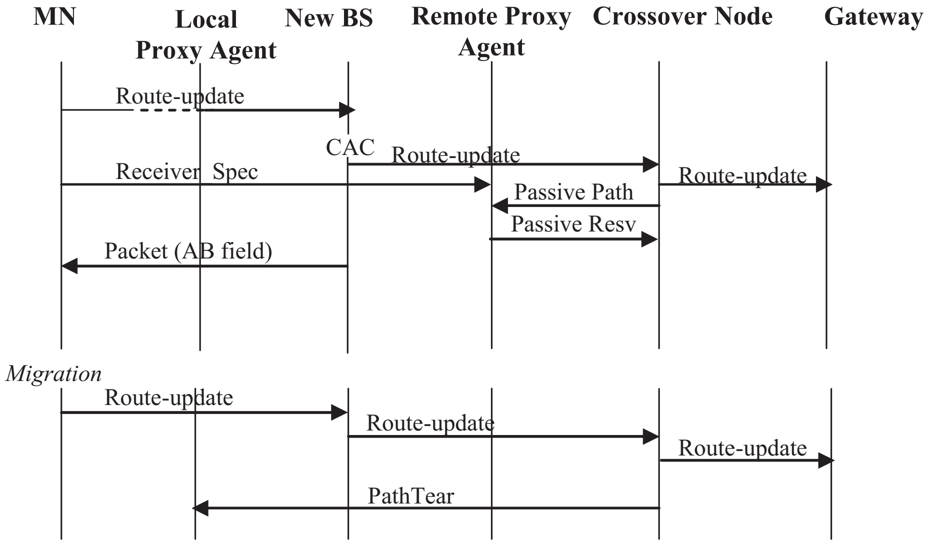


Figure 7.10: Signaling messages for Intra-domain Handover: the mobile node MN is a receiver

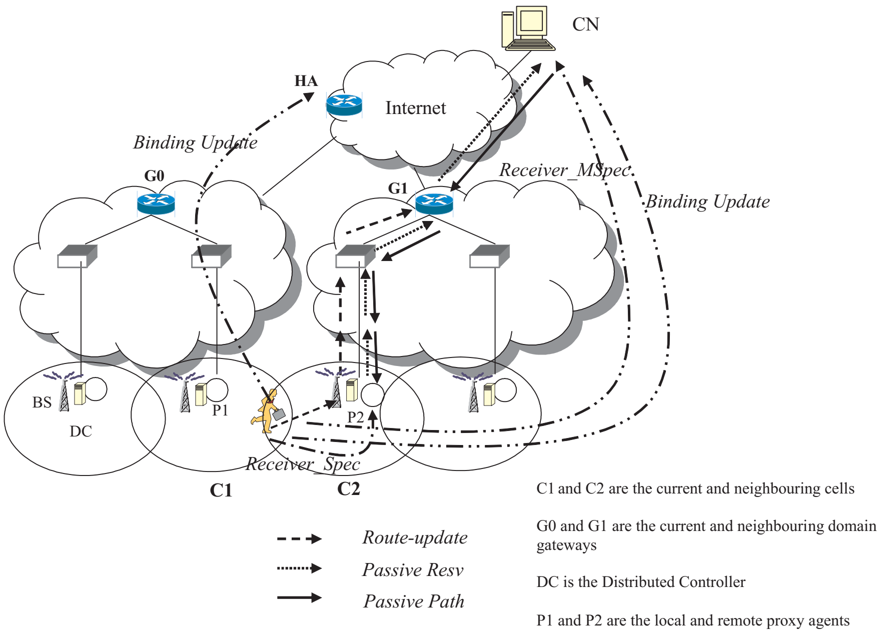

CN
HA
Internet
Binding Update
G0
G1
Receiver_MSpec
Binding Update
BS
DC
P1
P2
Receiver_Spec
C1
C2
Route-update
Passive Resv
Passive Path
C1 and C2 are the current and neighbouring cells
G0 and G1 are the current and neighbouring domain gateways
DC is the Distributed Controller
P1 and P2 are the local and remote proxy agents

Figure 7.11: Inter-domain Handover: the mobile node MN is a receiver

generated based on the Path state stored for the flow during previous RSVP message exchanges. That is after making sure that the ADSPEC and PHOP objects for each outgoing interface are updated, the RSVP module in the crossover node sends the Passive Path message along the forwarding path established freshly by the route-update message.

P1 sends a Passive Resv message to the crossover node. As a result, a passive reservation will be created in the neighboring cell.

4. The new base station sends a regular IP packet with an AB field to the MN and to the mobiles in the neighboring cell that should adapt their calls' bandwidth.

5. After migration the mobile node sends a route-update packet to the new Base Station with the S bit cleared. This route-update packet will remove all mappings in Route Cache except for the ones pointing to the new Base Station. We assume that the route-update packet can activate the passive reservation in the cell. Note that the crossover node must send a PathTear message to P0 in order to tear down the original active reservation in the cell C0.

Inter-domain handover When the MN is moving within the overlapping area of the cells C1 and C2 belonging to two different Cellular IP domains, an inter-domain handover is initiated and the following events occur (figure 7.11 and figure 7.12):

The first and second step are the same as the steps related to the intra-domain handover. Note that in this case, the crossover node is considered to be the gateway G1. The remaining

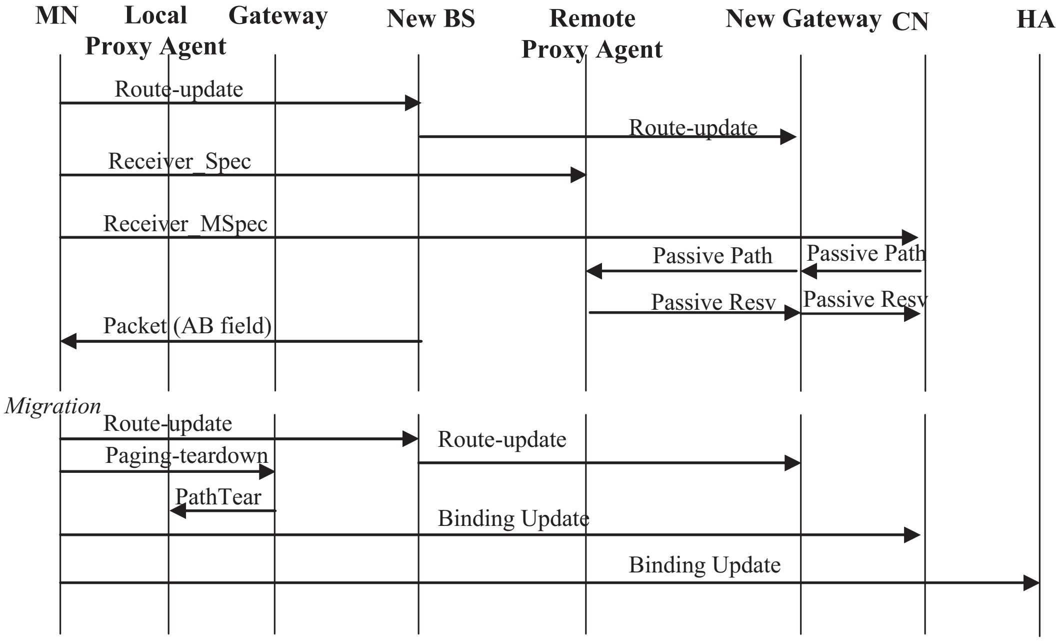


Figure 7.12: Signaling messages for Inter-domain Handover: the mobile node MN is a receiver

steps are as follows:

3. The MN sends a Receiver_MSpec (G1,P2) message to inform the CN that MN is visiting the cellular IP domain with G1 as gateway and P2 as proxy agent. By this message, the corresponding node CN initializes the building of passive resource reservation along the path from CN to P2 through G1. P2 and the CN will exchange a pair of Passive Path and Passive Resv. Hence a passive reservation will be then created in the neighboring cell.

4. The new base station sends a regular IP packet with an AB field to the MN and to the mobiles in the neighboring cell that should adapt their calls' bandwidth.

5. After migration the mobile node sends a route-update packet to the new Base Station with the S bit cleared. This route-update packet will remove all mappings in Route Cache except for the ones pointing to the new Base Station. We assume that the route-update packet can activate the passive reservation in the cell.

   The MN also sends a Paging-teardown to the old gateway G0 in order to clear the mappings associated to it in the Route and Paging caches. The gateway G0 will then issue a PathTear message to P1 to tear down the original active reservation in the cell C1.

   The mobile node should send a packet containing a Binding Update option to its Home Agent and the correspondent node notifying them about its new CoA.

7.7.4.2 Mobile node is a sender

Intra-domain handover When the MN is moving within the overlapping area of the cells C0 and C1 belonging to the same Cellular IP domain, an intra-domain Handover is initiated. The following events occur in the following order (figure 7.13 and figure 7.14):

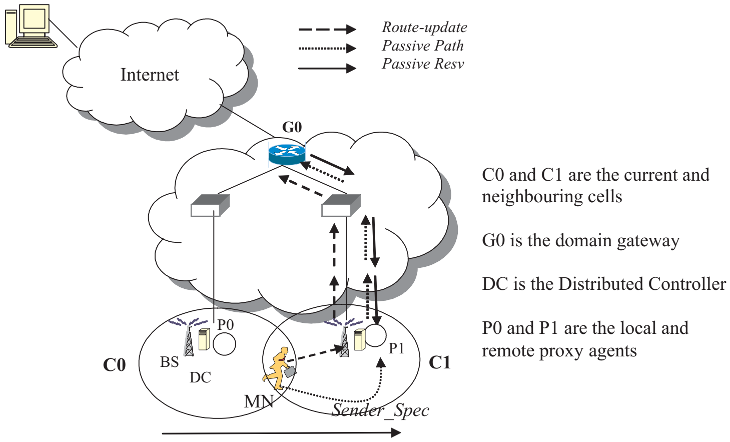

Route-update
Passive Path
Passive Resv
Internet
G0
C0 and C1 are the current and neighbouring cells
G0 is the domain gateway
DC is the Distributed Controller
P0 and P1 are the local and remote proxy agents
P0
BS
DC
C0
P1
C1
MN
Sender_Spec

Figure 7.13: Intra-domain Handover: the mobile node MN is a sender

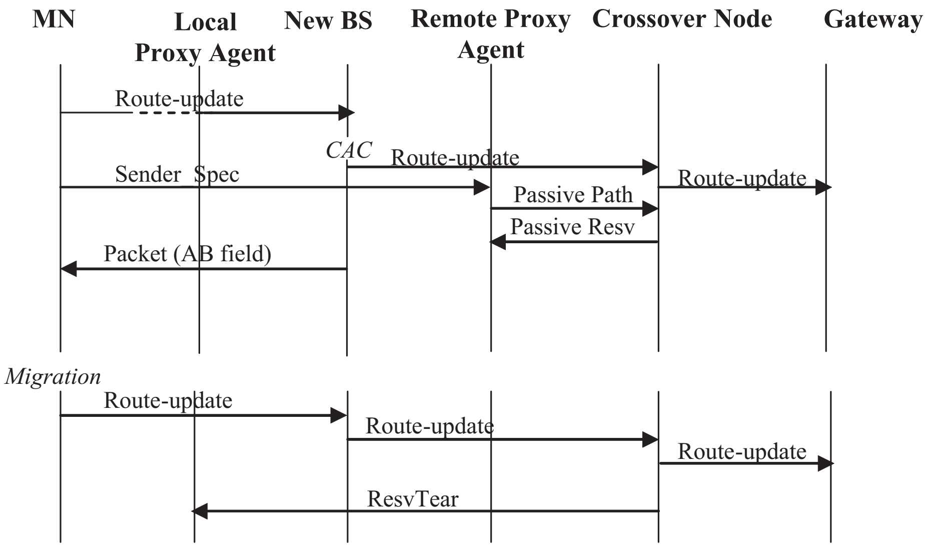


Figure 7.14: Signaling messages for Intra-domain Handover: the mobile node MN is a sender

1. The MN sends a route-update packet to the new base station located in the neighboring cell. The MN also sends a Sender_Spec to the remote proxy agent P1.

2. When the new base station receives the route-update packet sent by the MN, the DC in the neighboring cell performs the CAC and the bandwidth adaptation algorithms. If the HO request is accepted, then the new base station generates a semi-soft route-update packet. This semi-soft route-update packet creates new mappings in the Route and Paging Cache similarly to regular route-update packets. When the semi-soft route-update packet reaches the crossover node where the old and new path meet, the new mapping is added to the cache instead of replacing the old one.

3. On receiving the Sender_Spec message, the remote proxy agent P1 sends a Passive Path message to the crossover node. The crossover node will then issue a Passive Resv to P1. As a result, a passive reservation is created in the neighboring cell.

4. The new base station sends a regular IP packet with an AB field to the MN and to the mobiles in the neighboring cell that should adapt their calls' bandwidth.

5. After migration the mobile node sends a route-update packet to the new Base Station with the S bit cleared. This route-update packet will remove all mappings in Route Cache except for the ones pointing to the new Base Station. We assume that the route-update packet can activate the passive reservation in the cell. Note that the crossover node must send a ResvTear message to P0 in order to tear down the original active reservation in the cell C0.

Inter-domain handover The first and second step are the same as the steps related to the intra-domain handover. The remaining steps are as follows (figure 7.15 and figure 7.17):

3. On receiving the Sender_Spec message, the remote proxy agent P2 initializes the building of passive resource reservation along the path from P2 to the CN through G1. The remote proxy agent P2 and the CN will exchange a pair of Passive Path and Passive Resv. Hence a passive reservation will be then created in the neighboring cell.

4. The new base station sends a regular IP packet with an AB field to the MN and to the mobiles in the neighboring cell that should adapt their calls' bandwidth.

5. After migration the mobile node sends a route-update packet to the new Base Station with the S bit cleared. This route-update packet will remove all mappings in Route Cache except for the ones pointing to the new Base Station. We assume that the route-update packet can activate the passive reservation in the cell.

   The MN also sends a Paging-teardown to the old gateway G0 in order to clear the mappings associated to it in the Route and Paging caches. The gateway G0 will then issue a PathTear message to P1 to tear down the original active reservation in the cell C1.

   The mobile node should send a packet containing a Binding Update option to its Home Agent and the correspondent node notifying them about its new CoA.

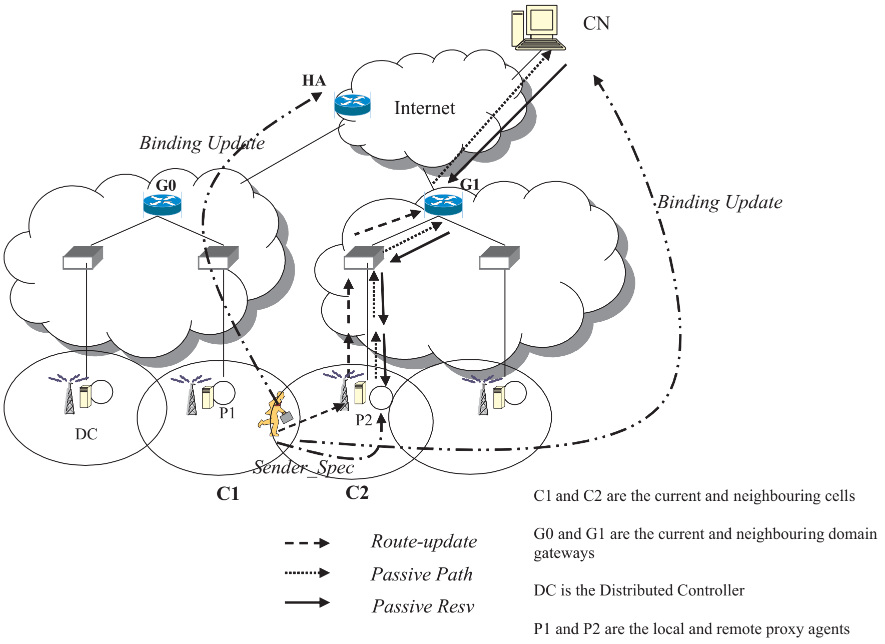

CN
HA
Internet
Binding Update
G0
G1
Binding Update
P1
P2
DC
Sender_Spec
C1
C2
Route-update
Passive Path
Passive Resv
C1 and C2 are the current and neighbouring cells
G0 and G1 are the current and neighbouring domain gateways
DC is the Distributed Controller
P1 and P2 are the local and remote proxy agents

Figure 7.15: Inter-domain Handover: the mobile node MN is a sender

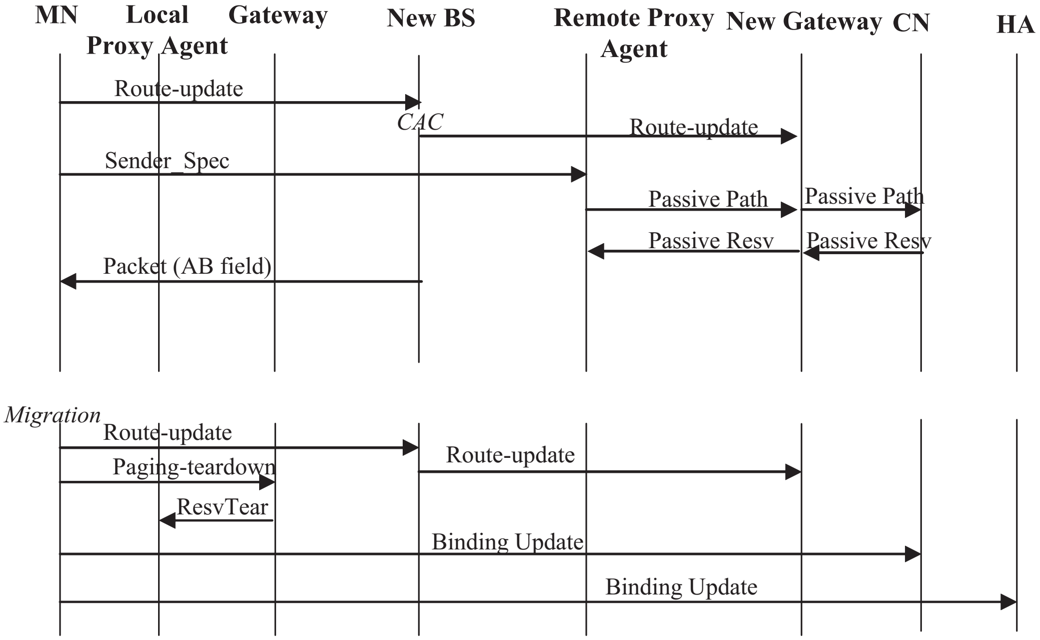


Figure 7.16: Signaling messages for Inter-domain Handover: the mobile node MN is a sender

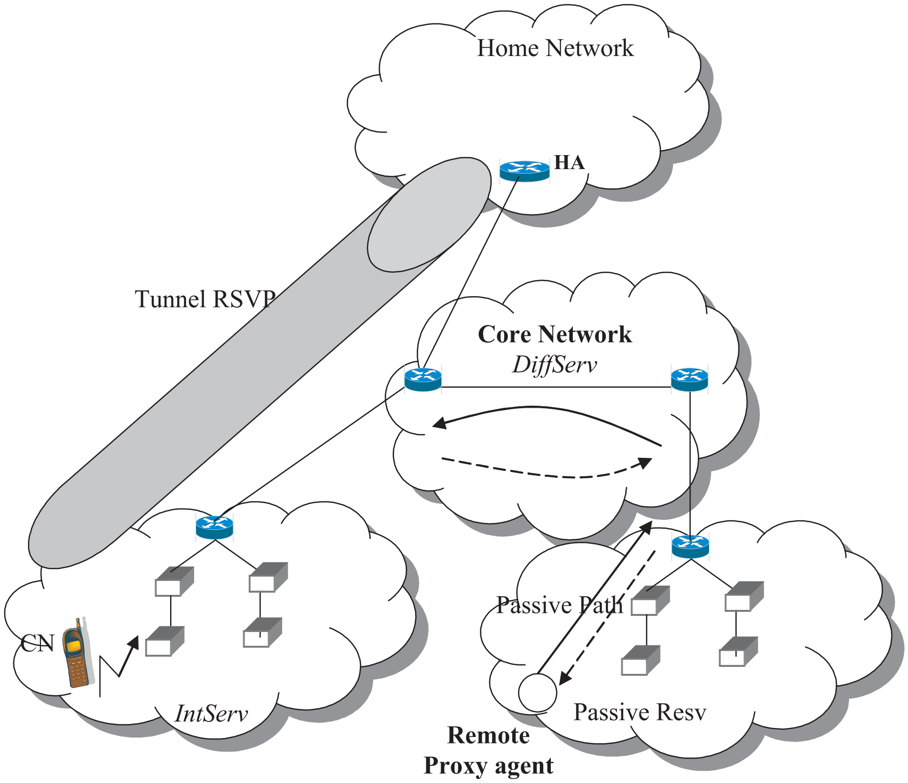


Figure 7.17: Signaling messages for Inter-domain Handover: Sender is mobile

7.7.5 Interaction of DiffServ and RSVP with Mobile IP

We proposed an end-to-end QoS approach with Cellular IPv6 implemented in the micro-mobility domains. As mentioned in chapter 3, the route optimization is used in Mobile IPv6. That is, when a corresponding node CN finds a cached binding for a destination address, the CN uses an IPv6 Routing Header to route the packet to the destination address. Alternatively, the packet can be sent encapsulated into an IPv6 tunnel header, if the CN does not find any cached entry for the destination address. In this case, the mobile node performs IPv6 decapsulation to extract the original IPv6 packet and then sends a Mobile IPv6 binding update to the packet sender.

When the IPv6 packets are encapsulated, RSVP will encounter the problems exposed in subsections 7.6.3 and 7.6.1. In order to solve the problems of the RSVP interaction with the Mobile IPv6, we propose the following solution.

Let us assume that the MN is a mobile sender. On the other hand, we suppose that the MN does not have a binding cache for the CN. In this case, the MN sends a Sender_Spec to the remote proxy agent (figure 7.17). The latter will issue a Passive Path to the new gateway which will send it to the CN. This message is sent through the core network which is DiffServ capable domain. The Passive Path message is mapped to the EF class of service within the DiffServ domain. The Passive Path message is then intercepted by the CN's HA. The Passive Path message will trigger *a RSVP tunnel* between the CN's HA and the CN, before being encapsulated by the HA. The encapsulated Path message will be mapped to the EF class of service in the DiffServ domain. Afterwards, the encapsulated Path message will arrive to the CN which decapsulates the packet.

7.8 Conclusion

The study in this chapter focuses on the design of an end-to-end QoS approach with Cellular IPv6 as the micro-mobility protocol in the micro-mobility domains. In a first step, we studied different types of CAC performed in a Cellular IPv6 network. A dynamic adaptive architecture was considered in order to enhance the performance. We showed the good performance achieved with the centralized/distributed CAC scheme, in terms of forced termination probability, new HA blocking probability and degradation parameters.

On the other hand, the distributed scheme has good performance and reduces the load and the processing in the gateway. The distributed scheme can be efficient and flexible in the case of highly bursty traffic. Thus, there is a trade-off between the centralized/distributed and the distributed scheme.

In a second step, we were interested in assuring an end-to-end QoS. Therefore, we started to study the problems of IntServ and DiffServ when applied to mobile environments. We then proposed to integrate IntServ in the Cellular IPv6 networks and DiffServ in the core network. The RSVP signaling needed to establish the intra-domain and inter-domain handovers was presented and discussed. We were concerned about minimizing the signaling load, especially in the radio interface. Therefore, we introduced the proxy agents that act on behalf of the mobile nodes. These proxy agents, attached to the base stations, detect and stop certain messages in order to reduce signaling overload in the network. On the other hand, we investigated the issue of the passive reservations that anticipate the mobile's movement.

The present chapter dealt with the QoS in the Cellular IP networks which are part of the fourth wireless generation networks. In the following chapter, we shall study the QoS aspect in IEEE WLANs. These networks are being presently deployed in local communication areas and are considered as a part of the fourth wireless generation networks as well. However, the IEEE WLANs lack of the QoS support. Thus, we shall define a novel mechanism within the MAC protocol as will be detailed in chapter 8.

Chapter 8

P3-DCF: Service Differentiation in IEEE 802.11 WLANs using Per-Packet Priorities

8.1 Introduction

In the recent years, much interest has been involved in the design of wireless networks for local area communication. Study group 802.11 was formed under IEEE Project 802 to recommend an international standard for wireless local area networks (WLAN) identified as IEEE 802.11 [LAN97]. The scope of the standard is the physical layer (PHY) and the medium access (MAC) sub-layer implementation.

Motivated by the growing use of multimedia applications, support for time-bounded services was also integrated in these IEEE 802.11 WLANs. This has been achieved by an extension of the basic medium access mechanism, the *Distributed Coordination Function* (DCF) using a centralized polling-based mechanism the *Point Coordination Function* (PCF). However, the proposed polling schemes based on the centralized decision are inefficient due to the transmission overheads.

In the present chapter, we enhance the MAC protocol defined within the IEEE 802.11 and introduce a novel mechanism that performs a service differentiation. In a first step, we study the DCF which is the basic medium access mechanism in the IEEE 802.11 WLANs. We afterwards present a survey of the research works dealing with the QoS enhancements in IEEE 802.11. Then, we propose our novel mechanism, called *P3-DCF*, employing the DCF as a fundamental access for prioritized service in IEEE 802.11 WLAN. An enhanced function, Per-Packet Priority (P3), integrated to the DCF mechanism establishes not only a per-flow differentiation but manages to schedule the packets with an Earliest Deadline First discipline as well.

The simulations carried out show that the proposed MAC mechanism satisfies the maximum tolerable latency of the real-time traffic and performs efficient flow differentiation. Moreover, the performance improvement is achieved without affecting the useful throughput.

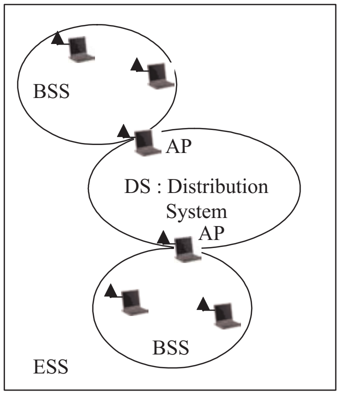


Figure 8.1: Infrastructure Network

8.2 IEEE 802.11 Architecture in the Standard

The Basic Service Set (BSS) is the fundamental building block of the IEEE 802.11 architecture. A BSS is defined as a group of stations that are under the direct control of a single coordination function, i.e, a DCF or PCF.

The geographical area covered by the BSS is known as the Basic Service Area (BSA) which is analogous to a cell in a cellular communication network (figure 8.1).

A single BSS can be used to form an ad-hoc network. An ad-hoc network is a deliberate grouping of stations into a single BSS for the purposes of inter-networked communications without the aid of an infrastructure network. In an ad-hoc network, any station can establish a direct communication session with any other station in the BSS, without the requirement of funneling traffic through a centralized Access Point (AP).

In contrast to the ad-hoc network, infrastructure networks are established to provide wireless users with specific services and range extension. Infrastructure networks in the context of IEEE 802.11 are established using APs. The AP is analogous to the base station in a cellular communication network. The AP links the wireless terminals to a Distribution System (DS), therefore extending their range to other BSSs via other APs. The DS can be any kind of fixed or wireless LAN. The whole system is then called Extended Service System (ESS).

8.3 MAC Sub-layer

The MAC sub-layer is responsible for the channel allocation procedures, protocol data unit (PDU) addressing, frame formatting, error checking, fragmentation and re-assembly. The transmission medium can operate in the contention mode exclusively, requiring all stations to contend for the channel for each packet transmitted. The medium can also alternate between the contention mode, known as the *Contention Period* (CP), and the *Contention Free Period* (CFP). During the CFP, the medium access is controlled by the AP, thereby eliminating the need for stations to contend for channel access.

8.3.1 Distributed Coordination Function

The DCF is the fundamental access method used to support asynchronous data transfer on a best effort basis. As identified in the specification, all stations must support the DCF.

Access to the medium in DCF The basic scheme for DCF is the Carrier Sense Multiple Access (CSMA). This protocol has two variants: Collision Detection (CSMA/CD) and Collision Avoidance (CSMA/CA). A collision can be caused by two or more stations using the same channel at the same time after waiting for the channel to become idle, or in wireless networks by two or more *hidden terminals* transmitting simultaneously. Hidden terminals are terminals which cannot hear each other [KT95].

CSMA/CD is used in Ethernet (IEEE 802.3) wired networks. With CSMA/CD, whenever a node detects that the signal it is transmitting is different from the one on the channel, it aborts transmission, saving useless collision time.

The DCF is based on the CSMA/CA protocol. The reason is that, even though the wireless LAN is a broadcast medium, the traditional CSMA/CD will not function properly because the station is unable to listen to the channel for collisions while transmitting, due to the big difference between transmitted and received power levels.

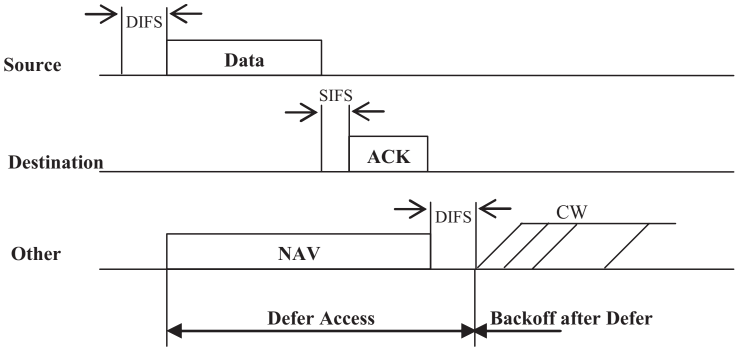


Figure 8.2: Transmission of MPDU without RTS/CTS

Carrier sensing In IEEE 802.11, carrier sensing is performed at both the physical layer and the MAC sub-layer. On the physical layer, the carrier sensing is referred to as the *physical carrier sensing*. Physical carrier sensing detects the presence of other IEEE 802.11 WLAN users and the activity in the channel via relative signal strength from other sources.

On the MAC sub-layer, carrier sensing is known as the *virtual carrier sensing*. Virtual carrier sensing is used by a source station to inform all other stations in the BSS of how long the channel will be used for the successful transmission of a MAC protocol data unit (MPDU). An MPDU is a complete data unit that is passed from the MAC sub-layer to the physical sub-layer. The MPDU contains header information, payload, and a 32-bit CRC. The source stations set the duration field in the MAC header of data frames, or in the Request to Send (RTS) and in the Clear to Send (CTS) control frames.

Stations detecting a duration field in a transmitted MPDU adjust their Network Allocation Vector (NAV) which indicates the amount of time that must elapse until the current transmission session is complete and then, the channel can be sampled again for idle status. The channel is marked busy if either the physical or the virtual carrier sensing mechanism indicates that the channel is busy.

Basic Transmission of a MPDU without RTS/CTS Contention services imply that each station with a MAC Service Data Unit (MSDU) at MAC level transport queued for transmission must contend for the channel, and once the MSDU is transmitted, must re-contend for the channel for all subsequent frames.

Priority access to the wireless medium is controlled through the use of *Inter-Frame Space* (IFS) time intervals between the transmission of frames. The IFS intervals are mandatory periods of idle time on the transmission medium. Three IFS intervals are specified in the standard: *Short-IFS* (SIFS), *Point Coordination Function-IFS* (PIFS) and *Distributed Coordination Function* (DIFS). The SIFS interval is the smallest IFS, followed by PIFS, followed by DIFS.

Figure 8.2 is a timing diagram illustrating the successful transmission of a data frame. When the data frame is transmitted, the duration field of the frame is used to let all the stations in the BSS know how long the medium will be busy. All stations hearing the data frame, adjust their NAV based on the duration field value which includes the SIFS interval and the acknowledgment frame following the data frame.

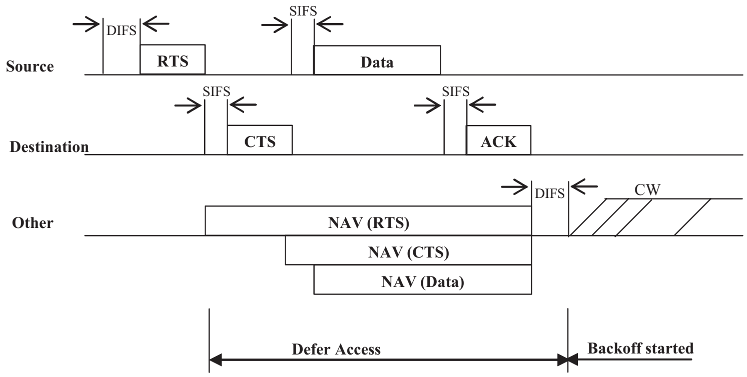


Figure 8.3: Transmission of MPDU with RTS/CTS

Backoff Procedure The collision avoidance portion of CSMA/CA is performed through a random backoff procedure. If a station with a frame to transmit initially senses the channel to be busy, then the station waits until the channel becomes idle for a DIFS period and then computes a random *backoff time*. For IEEE 802.11, time is slotted in time periods called Slot_times. The random backoff time is an integer value that corresponds to a number of Slot_times.

Initially, the station computes a backoff time uniformly in the range $0 - 7$ Slot_times. The backoff time of each station is decreased as long as the channel is idle during the so-called *contention window CW*. When the channel is busy and the backoff time has not reached zero, the station *freezes* its backoff time. When the backoff time reaches zero, the station transmits its frame. If two or more stations decrement to zero at the same time, then a collision will occur and each station will have to generate a new backoff time in the range $0 - 15$ Slot_times.

For each retransmission attempt, the backoff grows exponentially as:

$$T_{backoff} = \lfloor 2^{2+i} * rand() \rfloor * Slot\_time \tag{8.1}$$

where:

- i is the number of consecutive times a station attempts to send a MPDU.
- $rand()$ is a random function with a uniform distribution in $(0, 1)$.
- $\lfloor x \rfloor$ represents the largest integer less than or equal to x.

The advantage of this channel access method is that it promotes fairness among stations. Fairness is maintained because each station must re-contend for the channel after every transmission of a frame. All stations have equal probability of gaining access to the channel after each DIFS interval.

Other references ([BCV01], [Bia00]) compute the backoff time such that:

$$T_{backoff} = Rand(0, CW) * Slot\_time \tag{8.2}$$

With the above formula, after each unsuccessful transmission attempt, the Contention Window CW is doubled until a predefined maximum (CW_{max}) is reached. The starting value of CW is CW_{min}.

Transmission of a MPDU using RTS/CTS Since a source station in a BSS cannot hear simultaneously, when a collision occurs (caused by a hidden terminal for example), the source continues transmitting the complete MPDU. If the MPDU is large, the bandwidth is wasted due to a corrupted MPDU.

To solve the hidden terminal problem, an optional RTS/CTS (Request To Send/Clear To Send) scheme is used in addition to the previous basic scheme as shown in figure 8.3. RTS and CTS control frames can be used by a station to reserve a channel bandwidth prior to the transmission of an MPDU so that the bandwidth waste due to collisions is minimized.

When hidden nodes exist, a collision of RTS frames (20 bytes) is less severe and less probable than a collision of data frames (up to 2346 bytes). In fact, a station may sense the channel to be idle and start a transmission, while another transmission from a hidden terminal is taking place, causing collision. Therefore, in hidden terminal scenarios, collision probability is proportional to packet sizes, and using short RTS/CTS reduces collision probability and the bandwidth waste during collisions. The destination replies with a CTS if it is ready to receive and the channel is then reserved for the packet duration. When the source receives the CTS, it starts transmitting its frame, being sure that the channel is reserved for itself during all frame duration. All other stations in the BSS update their Network Allocation Vector whenever they hear an RTS, a CTS or data frame. Figure 8.3 illustrates the transmission of an MPDU using the RTS/CTS mechanism.

Time-bounded services typically support applications such as packetized video or voice that must be maintained with a specified delay. As defined in the standard [LAN97], DCF does not guarantee delay to stations supporting time-bounded services. Therefore, research works have proposed enhancement to the DCF mechanism as presented in the next section.

8.4 Service Differentiation in IEEE 802.11

Several approaches compete for providing QoS enhancements in IEEE 802.11: The IEEE 802.11 working group has provided basic real-time support by introducing the PCF. The PCF is an optional capability, which is connection-oriented, and provides contention-free services enabling polled stations to transmit without contending for the channel. The polling function is performed by the AP within each BSS.

This centralized polling scheme might be enhanced by using improved scheduling or signaling schemes. In [ST01], the authors propose two types of priority scheduling schemes according to whether priority control is only supported by the AP or by both the AP and the wireless terminals. They show that priority schemes support larger number of multimedia terminals than the non-priority scheme. However the data throughput performance with the two priority schemes is worse than with the non-priority scheme.

The authors in [KW01] compared the two mechanisms DCF and PCF by simulation. They simulated a DCF scheme coupled with a local scheduling policy on the mobile nodes as well as the access point, and a normal DCF without any scheduling policy. The authors found that PCF reduces the mean channel access delay by avoiding contention phases during CFP and reducing contention in CP. In addition, it offers a very low variance. The authors concluded that PCF shows better performance than DCF outlining the fact that a centralized access control scheme should be used at high data rates. However, the polling schemes based on the centralized decision have inefficiency due to the transmission overheads.

Some proposals introduce service differentiation in the MAC layer using the DCF mechanism.

The authors in [BCV01] extended the Distributed Coordination Function by integrating two distributed estimation algorithms: the Virtual MAC (VMAC) and the Virtual Source (VS). The VMAC and the VS algorithms passively estimate whether the channel can support new service demands taking into account both local conditions and interference caused by external effects. Based on the service quality estimations obtained from the virtual monitoring algorithms, mobile nodes and base stations determine whether a new session with a particular service level requirement should be admitted. The authors showed that if all nodes use passive monitoring and base their admission decisions accordingly, a globally stable state can be maintained without the need for complex centralized radio resource management.

In [KSK01], the authors propose a new access control employing a proposed "DCF/SC" instead of PCF, and DCF as two fundamental accesses for priority services and apply three kinds of prioritized services (premium, assured and best-effort). The proposed DCF/SC has a shorter contention window than that of the original DCF. The proposed mechanism satisfies the maximum tolerance bounds of the traffic sources and enhances the channel utilization. Nevertheless, it suffers from the fact that it implies changes in the standards protocol behavior.

In [AC03], the main issue is the QoS adapted tuning of the DCF backoff mechanism, and of some important parameters relative to DCF. In order to introduce a service differentiation in the DCF operation, the authors in [AC03] assign priorities to flows. These priorities can then be applied to flows using one of these possible policies:

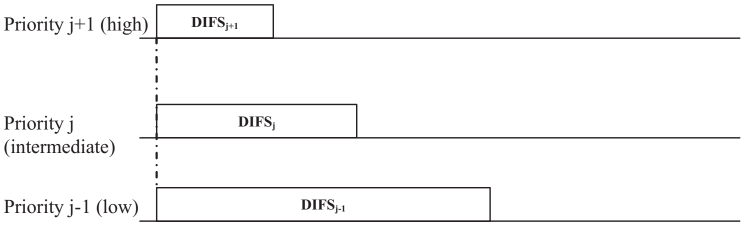


Figure 8.4: DIFS Differentiation Mechanism

1. CW_{min} and CW_{max} differentiation mechanism: different levels of service (i.e., delay and loss) can be achieved by setting different values of CW_{min} and CW_{max} for the different classes of service.

2. Backoff increase function: as stated above, each time the transmission fails, CW is doubled, i.e. new CW =(old CW)*2. Considering that the only configurable term is 2, an attempt to introduce priority is to replace it by P_j, where P_j is a priority factor assigned to the service class j. The higher the priority factor is, the larger is the backoff range, the lower is the chance to first access the channel, the lower is the throughput.

3. DIFS differentiation mechanism: another attempt to perform service differentiation among flows is to use different DIFS values for the different classes of service. If a class j has higher priority than the class $j-1$, the $DIFS_j$ value corresponding to the class j must be less than $DIFS_{j-1}$ corresponding to the class $j-1$ (figure 8.4). Note that, in the standard [LAN97], the DIFS value is initialized for all packets as follows:

$$DIFS = SIFS + 2 * Slot\_time. \tag{8.3}$$

And with the DIFS differentiation mechanism, the DIFS value for the j^{th} class is defined by the following equation:

$$DIFS_j = SIFS + nbSlotDIFS_j * Slot\_time. \tag{8.4}$$

where $nbSlotDIFS_j$ is a differentiation parameter that allows to compute $DIFS_j$ as defined above.

The authors in [AC03] demonstrate that the DIFS differentiation mechanism is the best way to differentiate the flows. Starting from this result, we design a new differentiation mechanism that deals with the queuing delays, experienced by each packet, in the computation of the DIFS value. To this end, we define a per-packet priority function as explained in the next section.

8.5 Per-Packet Priority: P3-DCF

Our proposed mechanism P3-DCF ensures that all packets sent by a mobile host are differentiated. Note that the differentiation is effective among packets sent by other mobile hosts as well.

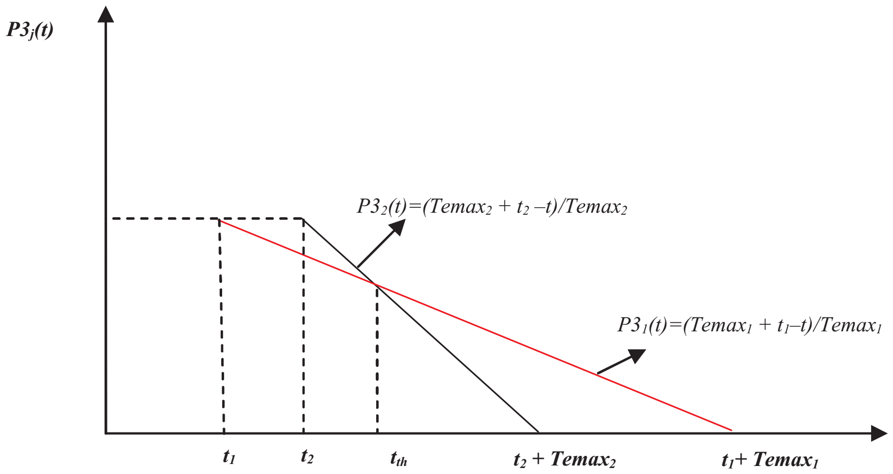


Figure 8.5: Per-packet Priority Function $P3_j(t)$

This is achieved by adding the *Per-packet priority* function that computes the expiration delay of a packet and that sends the urgent packets accordingly.

8.5.1 Per-packet priority function

Let us define a traffic class j by the set $(Temax_j, DIFS_j^{min}, DIFS_j^{max})$, where:

- $DIFS_j^{min}$ (respectively $DIFS_j^{max}$) represents the minimum (respectively the maximum) value of DIFS corresponding to the j^{th} class of service.
- $Temax_j$ is the maximum waiting delay to be experienced by a packet belonging to the class j.

Let us suppose that a packet belonging to the j^{th} class of service arrives at time τ. A Per-packet priority function $P3_j(t)$ is computed and assigned to this packet at the current time t, such that:

$$P3_j(t) = (Temax_j + \tau - t)/Temax_j \tag{8.5}$$

We assume that the packet with the lowest value of $P3_j(t)$ has the highest priority. Note that $P3_j(t) \in [0, \ldots, 1]$. Whenever $P3_j(t) = 0$, the packet is dropped.

The proposed function $P3_j(t)$ can be seen as the implementation of the Earliest Deadline First algorithm (EDF) and the proposed system is a *time-dependent priority* system [Kle76].

Figure 8.5 shows an example of the manner in which the priority is handled between two packets belonging to two different classes of service. Specifically, at time t_1, a packet p_{11} from class 1 arrives and has a deadline equal to $Temax_1$. At time t_2, a different packet p_{21} enters from a higher priority class of service 2; that is $Temax_2 < Temax_1$. Note that τ in the formula 8.5 is equal respectively to $t1$ and $t2$ for the packets p_{11} and p_{21}.

When the medium is idle, the packet with the highest priority (i.e. with the lowest value of $P3_j(t)$) is chosen. Thus, in our example, the first packet p_{11} will be chosen in preference to the second packet p_{21} if the medium is free at any time between t_1 and t_{th}, where t_{th} corresponds to the time at which the function $P3_j(t)$ has the same value for both packets. The first packet is chosen in spite of the fact that it is from a lower priority class. However, for any time after t_{th}, the second packet will be chosen in preference to the first.

8.5.2 DIFS differentiation using the Per-packet priority function

As stated above, the DIFS differentiation has been found in the literature to be the best way to differentiate the flows. Our mechanism enhances the DIFS differentiation with the integration of the Per-packet priority function into to the DIFS computation.

More specifically, the MAC sub-layer, when processing a packet belonging to a class of service j at time t, computes the $DIFS_j(t)$ value as follows:

$$\begin{aligned} DIFS_j(t) &= SIFS + (nbSlotDIFS_j^{min} + \\ &\lfloor (nbSlotDIFS_jmax - nbSlotDIFS_jmin) * P3_j(t) \rfloor) * Slot\_time \end{aligned} \tag{8.6}$$

Where

- t is the time where the MAC sub-layer begins processing the packet.
- $DIFS_j(t)$ is the DIFS value computed at time t for the current packet belonging to the j^{th} class of service.
- $\lfloor x \rfloor$ represents the largest integer less than or equal to x.
- $(nbSlotDIFS_j^{max}, nbSlotDIFS_j^{min})$ are two parameters allowing us to compute $DIFS_j^{min}$ and $DIFS_j^{max}$ as follows:

$$DIFS_j^{min} = SIFS + nbSlotDIFS_j^{min} * Slot\_time \tag{8.7}$$

$$DIFS_j^{max} = SIFS + nbSlotDIFS_j^{max} * Slot\_time \tag{8.8}$$

The computation of the $DIFS_j(t)$ value shows that each packet is assigned a priority that is time-dependent. In this way, our novel mechanism P3-DCF establishes not only a per-flow differentiation but manages to schedule the packets with an Earliest Deadline First discipline as well. Consequently, the QoS metrics of the real-time traffic are satisfied, and the per-flow differentiation is established as will be shown in the simulation results.

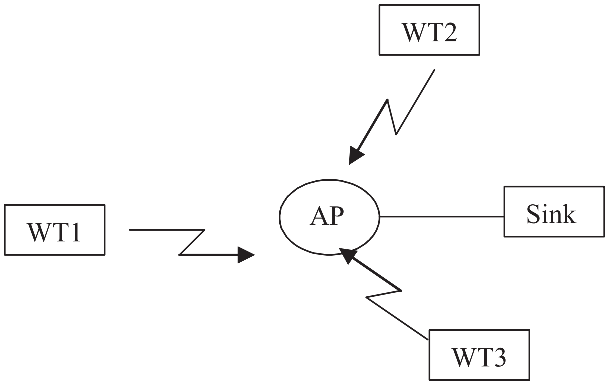


Figure 8.6: Simulation Network Topology

8.6 Simulation

In order to study the performance of the proposed mechanism, some simulations were carried out using ns-2.1b9 [Net]. The topology of the simulation network is rather simple: three wireless terminals (WT1, WT2 and WT3) are uniformly distributed around an access point and send three competing flows to an AP (figure 8.6). The useful capacity of the 1 Mbps wireless link is equal to 0.8 Mbps. Simulation time is 250s.

In order to evaluate the performance of our mechanism, the DIFS differentiation mechanism was simulated for comparison purposes.

8.6.1 UDP/CBR Traffic

In a first step, we consider that the three terminals send three UDP/CBR flows namely; CBR1, CBR2 and CBR3. We gave CBR1 higher priority than CBR2 which has in turn a higher priority than CBR3. CBR1, CBR2 and CBR3 started sending packets respectively at seconds 50, 100 and 150. Each flow overloads the link with 2312-byte packets every 0.02s. The simulation parameters relative to the P3-DCF and the DIFS differentiation mechanism are shown in table 8.1 and 8.2.

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
CBR1	$50\mu s/130\mu s$	$150ms$
CBR2	$130\mu s/210\mu s$	$250ms$
CBR3	$210\mu s/290\mu s$	$350ms$

Table 8.1: Simulation parameters for the P3-DCF Mechanism

Figures 8.7 and 8.8 plot the mean end-to-end delay of the three flows. The curves in figure 8.7 show that the flows differentiation is better with the P3-DCF mechanism than with the simple DIFS differentiation mechanism (figure 8.8). On the other hand, the delays obtained with our enhanced mechanism are smaller than those experienced with the DIFS differentiation mechanism. We expected this since in our proposed mechanism, we try to keep all the packet waiting delay less than $Temax_j$. Our mechanism reduces delays by dropping packets that arrived

	DIFS
Wireless Terminal WT1	$50\mu s$
Wireless Terminal WT2	$130\mu s$
Wireless Terminal WT3	$210\mu s$

Table 8.2: Simulation Parameters for the DIFS Differentiation Mechanism

at their transmission deadline, rather than having their backoff time increased. Furthermore, one can see that the delay variation (jitter) is smaller when using our mechanism.

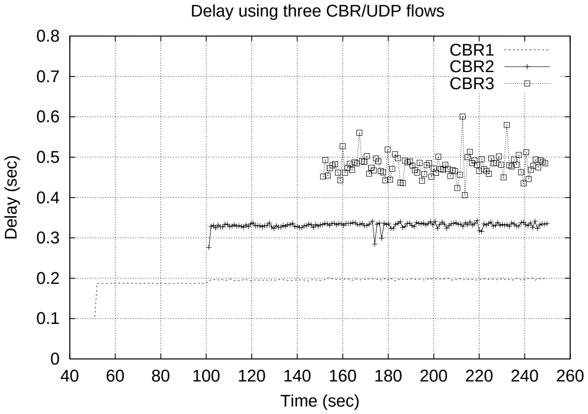


Figure 8.7: Mean Delay Using UDP with the P3-DCF Mechanism

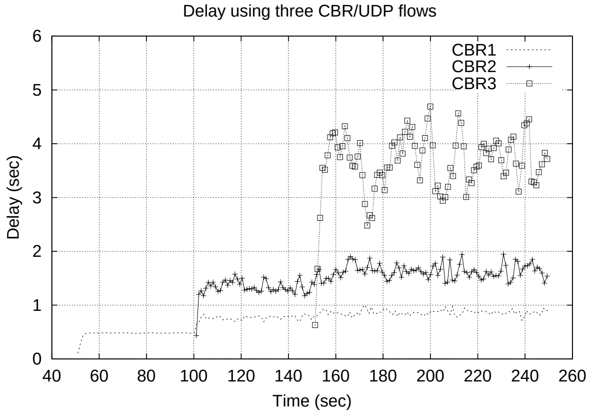


Figure 8.8: Mean Delay Using UDP with the DIFS Differentiation Mechanism

The loss ratio and throughput do not change from those in the DIFS differentiation mechanism (figures 8.9, 8.10, 8.11 and 8.12). In fact, in the DIFS differentiation mechanism packet loss occurs when the transmission attempt limit is reached. Let us denote the packet loss probability of the DIFS differentiation mechanism by P_{Loss}. In the P3-DCF mechanism, the packet loss probability is equal to $P_{LossTemax} + P_{LossBackoff}$, where $P_{LossTemax}$ is the probability of losing a packet whenever its deadline, $Temax_j$, is reached. $P_{LossBackoff}$ is the probability a packet arriving at its transmission attempt limit is attained. With P3-DCF, $P_{LossTemax}$ increases and $P_{LossBackoff}$ decreases, such that $P_{LossTemax} + P_{LossBackoff}$ remains equal to P_{Loss}. This is because the link is overloaded.

8.6.2 TCP/FTP Traffic

Other interesting simulation results were also obtained by replacing the UDP by the TCP transport layer. In this case, the three terminals send three TCP/FTP traffics namely; FTP1, FTP2 and FTP3. FTP1 is supposed to have more priority than FTP2 which has in turn more priority than FTP3. FTP1, FTP2 and FTP3 start sending packets respectively at seconds 50, 100 and 150. Each single traffic sends 1100 bytes FTP packets.

The simulations parameters relative to the P3-DCF and the DIFS differentiation mechanism are given in tables 8.3 and 8.2.

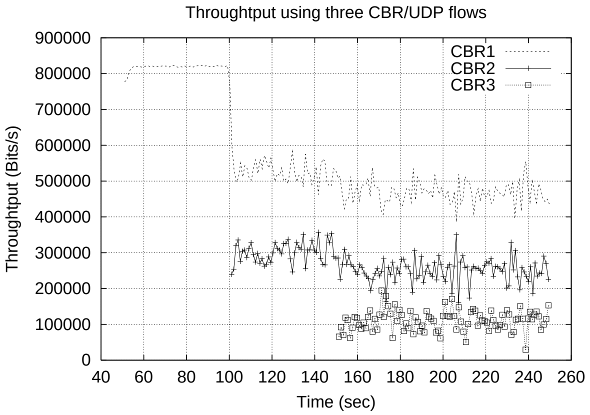


Figure 8.9: Throughput Using UDP with the DIFS Differentiation Mechanism

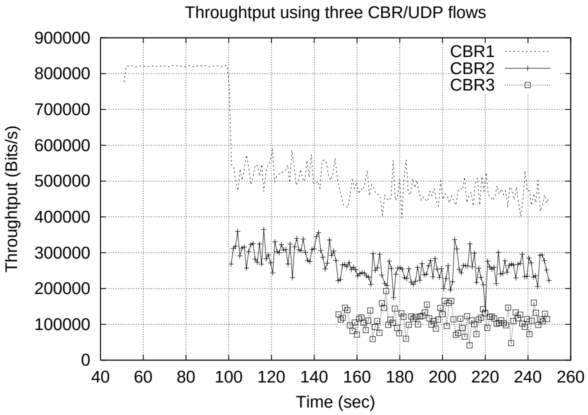


Figure 8.10: Throughput Using UDP with the P3-DCF Mechanism

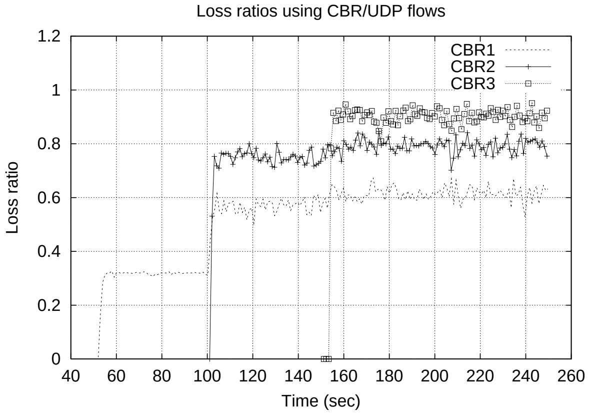


Figure 8.11: Loss Ratio Using UDP with the DIFS Differentiation Mechanism

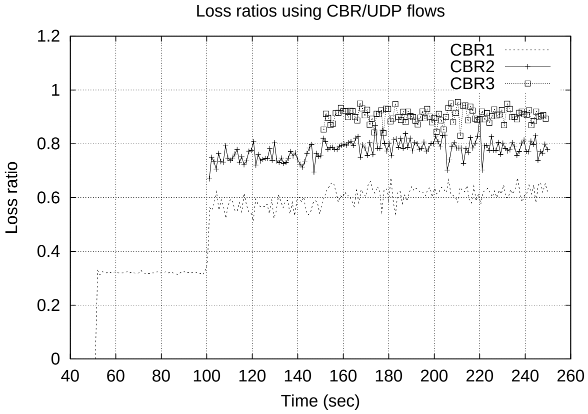


Figure 8.12: Loss Ratio Using UDP with the P3-DCF Mechanism

As shown in Figures 8.13, 8.14 and 8.15, the flow FTP1 which has the highest priority level (i.e. the smallest values for the set $(DIFS_j^{min}/DIFS_j^{max}/Temax_j)$ experiences bad performance: the $Temax$ is so small that a high number of packets is dropped by our mechanism and hence, retransmitted by TCP. Thus, the FTP1 flow enters often in the TCP congestion avoidance phase (Figure 8.14). Moreover, this situation is even worst when the second flow FTP2 begins sending packets: one can see that the flow prioritization between these two flows does not exist (Figures 8.13, 8.14 and 8.15).

When the third flow FTP3 starts, the packet loss of the first flow becomes so high that it stops sending packets after a few seconds. This result remains the same when $Temax$ values are increased (e.g. Table 8.4). In fact, high values of $Temax$ for a low priority flow give more chance to its packets to wait and to be transmitted. Then, the packets belonging to the highest priority flow are most of the time dropped and retransmitted by TCP [1]. This fact incurs overhead due to the TCP-level retransmissions that penalize the highest priority flow and that leads to bad

[1] This result is only valid when giving small $Temax$ values to flows with high priority

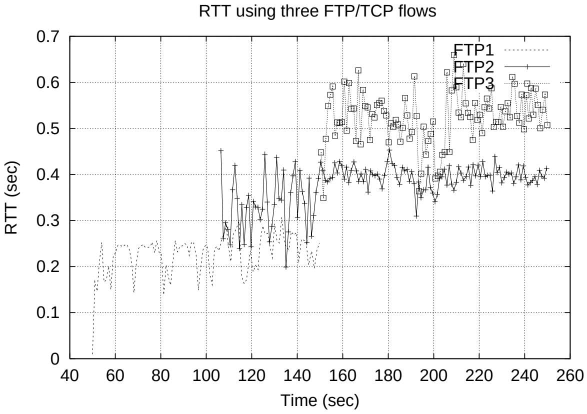


Figure 8.13: RTT Using TCP with the P3-DCF Mechanism (small $Temax$ values)

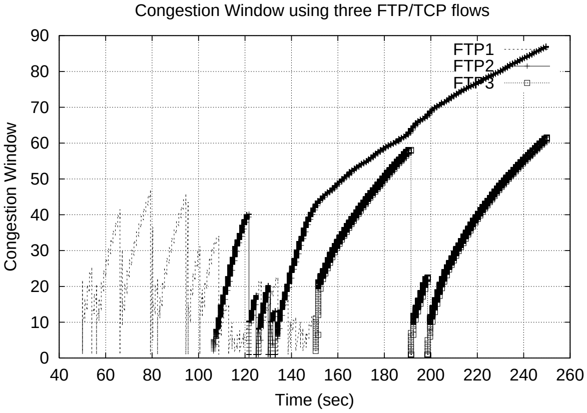


Figure 8.14: Congestion Window Using TCP with the P3-DCF Mechanism (small $Temax$ values)

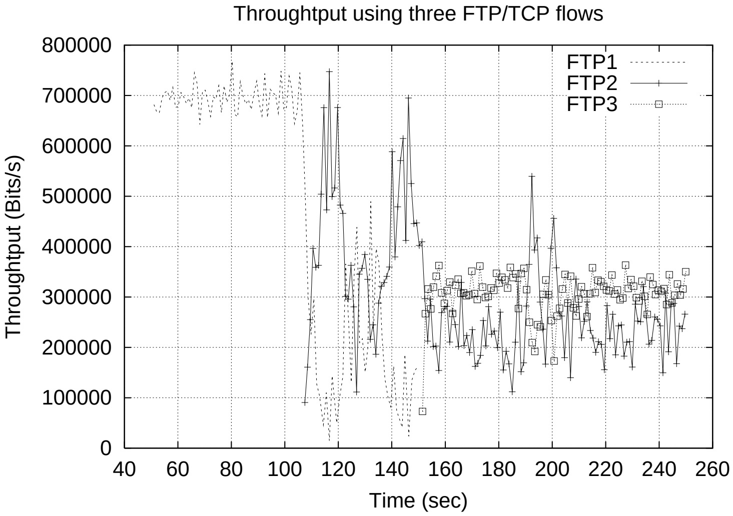


Figure 8.15: Throughput Using TCP with the P3-DCF Mechanism (small $Temax$ values)

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
TCP1	$50\mu s/130\mu s$	$150ms$
TCP2	$130\mu s/210\mu s$	$250ms$
TCP3	$210\mu s/290\mu s$	$350ms$

Table 8.3: Simulation Parameters Using TCP for the P3-DCF Mechanism (small $Temax$ values)

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
TCP1	$50\mu s/130\mu s$	$300ms$
TCP2	$130\mu s/210\mu s$	$500ms$
TCP3	$210\mu s/290\mu s$	$700ms$

Table 8.4: Simulation Parameters Using TCP for the P3-DCF Mechanism (high $Temax$ values)

service differentiation. At this point, we can state that the $Temax$ value must be fine tuned in order to establish the prioritization between the competing flows.

As a result, the best way to achieve service differentiation among TCP flows is to assign the same values of $Temax$ for all TCP flows (Table 8.5). Following this result, new simulations were carried out using parameters depicted in Tables 8.5 and 8.2. The results are compared with those obtained when using the DIFS differentiation mechanism.

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
TCP1	$50\mu s/130\mu s$	$375ms$
TCP2	$130\mu s/210\mu s$	$375ms$
TCP3	$210\mu s/290\mu s$	$375ms$

Table 8.5: Simulation Parameters Using TCP for the P3-DCF Mechanism

With the same value of $Temax$, the performance of the P3-DCF is improved. Figures 8.16 to 8.21 show better service differentiation with our mechanism than with the DIFS differentiation mechanism. In fact, with P3-DCF, the RTT of the highest priority flows are smaller than those obtained when using the DIFS differentiation mechanism. Moreover, the delay variation is decreased with P3-DCF.

When analyzing the congestion window evolution curves plotted in figure 8.19, we notice that the lowest priority flow FTP3 enters most of the time in the slow start phase of TCP. Thus, this limits its throughput compared to the throughput of other flows leading to a better service differentiation (Figures 8.19 and 8.21).

With the DIFS differentiation mechanism, the service differentiation between the three competing TCP flows is not visible as is the case when transmitting UDP flows. Indeed, we observe no TCP dropped packets due to the TCP adaptability: when the sender requests to transmit and the channel is idle, no dropping is observed as long as the traffic is adapted to the offered

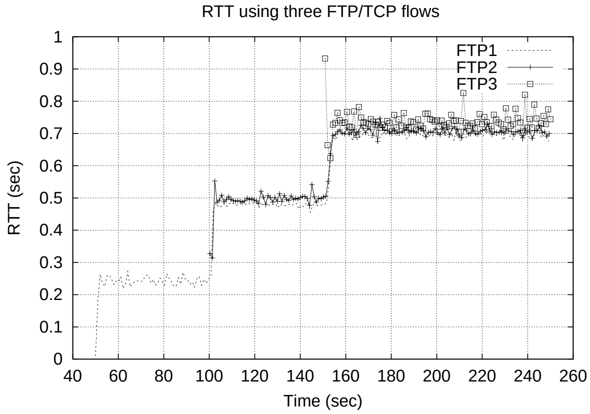


Figure 8.16: RTT Using TCP with the DIFS Differentiation Mechanism

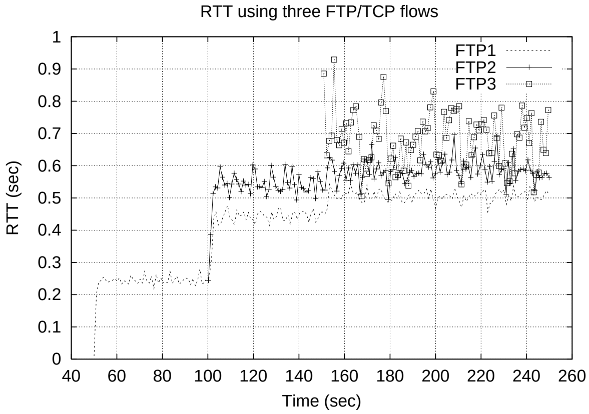


Figure 8.17: RTT Using TCP with the P3-DCF Mechanism

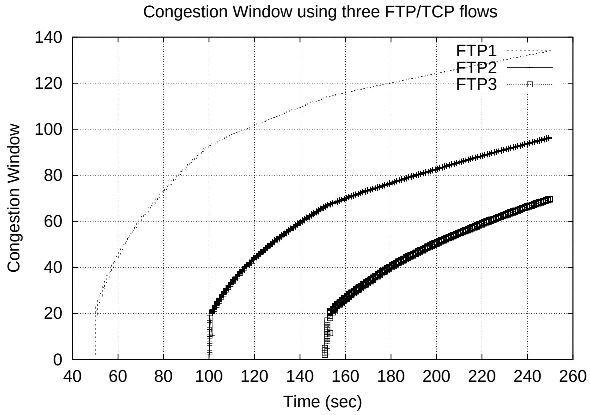


Figure 8.18: Congestion Window Using TCP with the DIFS Differentiation Mechanism

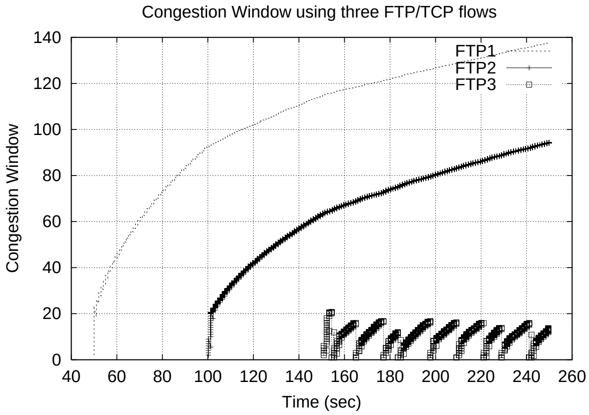


Figure 8.19: Congestion Window Using TCP with the P3-DCF Mechanism

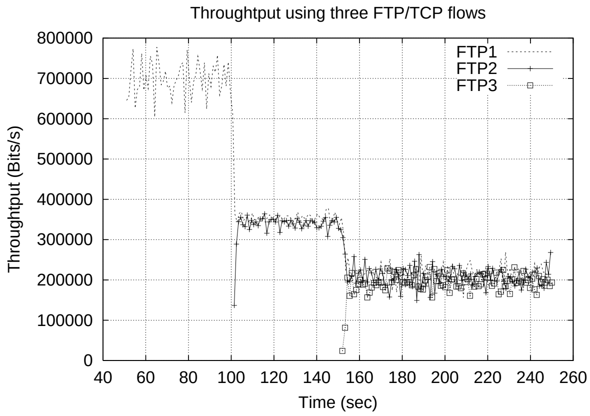


Figure 8.20: Throughput Using TCP with the DIFS Differentiation Mechanism

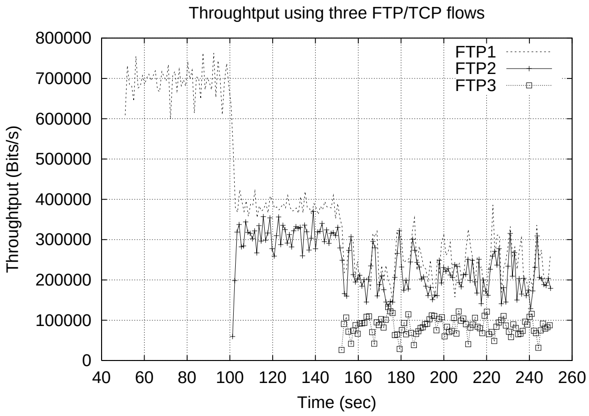


Figure 8.21: Throughput Using TCP with the P3-DCF Mechanism

throughput, which is the case of TCP. At each new period, more congestion occurs and the general slope of the congestion window evolution decreases (Figure 8.18).

However, the congestion window never decreases during the simulation time. At each TCP-ACK packet arrival, the congestion window increases, and it never decreases because TCP never times out for a TCP-ACK reception: dropped RTSs, for TCP-ACK and data packets, are retransmitted by the MAC sub-layer much faster than the TCP timeout. Thus, packet dropping due to buffer overflow at the sender is avoided with TCP and this attenuates the service differentiation with the DIFS differentiation mechanism as compared with P3-DCF.

With our mechanism, packets belonging to the lowest priority flow are dropped because of the time-out of the $Temax$. This slows down the packets transmission of the flow in question. Consequently, priority flows with high priority profit from the throughput non used by lower priority flows (Figure 8.21). We can conclude that there is no bandwidth loss.

8.6.3 Mixed UDP/CBR and TCP/FTP traffics

As we have seen before, P3-DCF improves the performance when the terminals send TCP traffic or UDP traffic separately. In this section, we analyze the impact of P3-DCF on a mix of TCP and UDP flows. As UDP is often used by real-time traffic, UDP has priority over TCP. Thus, we investigate if the QoS of UDP flow is maintained in the presence of competing TCP flows.

To this end, we have simulated two different scenarios. The first one comprises one TCP/FTP flow and two UDP/CBR flows. The second scenario comprises two TCP/FTP flows and one UDP/CBR flow.

8.6.3.1 One flow TCP and Two flows UDP

The generated flows are FTP1, CBR2 and CBR3. FTP1 has lower priority than CBR2 which in turn has lower priority than CBR3. FTP1, CBR2 and CBR3 start sending packets respectively at seconds 50, 100 and 150. The simulation parameters are given in Table 8.6 and Table 8.2.

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
CBR3	$50\mu s/130\mu s$	$150ms$
CBR2	$130\mu s/210\mu s$	$250ms$
FTP1	$210\mu s/290\mu s$	$1s$

Table 8.6: Simulation Parameters for the P3-DCF Mechanism

Figures 8.22 and 8.23 plot the mean end-to-end delay of the three generated flows with both mechanisms. The mean FTP delay increases each time a CBR flow enters the system. This shows the flow differentiation achieved with both mechanisms. In addition, one can also see that our mechanism decreases the mean delay of the CBR3 flow. Once again, we see that the delay variation of the CBR flows is reduced.

We explain this result by analyzing the congestion window (cwnd) evolution retrieved with both mechanisms (Figures 8.24, 8.25). Figure 8.25 shows that with P3-DCF, the slope of the congestion window decreases when the first UDP flow CBR2 enters to the system at second 100s. When the second UDP flow CBR3 arrives, the TCP flow FTP1 enters the slow-start phase. Afterwards, FTP1 experiences continuous slow start phases. Indeed, UDP flows have higher priority than TCP flows. In other words, UDP packets have more chance to be sent than TCP packets.

Figure 8.26 shows that the throughput of CBR flows is slightly smaller than that of the P3-DCF mechanism (figure 8.27). The loss ratio obtained with the P3-DCF is smaller than that of the DIFS differentiation mechanism (figure 8.28, figure 8.29). This is done without leading TCP flows to stop their transmission.

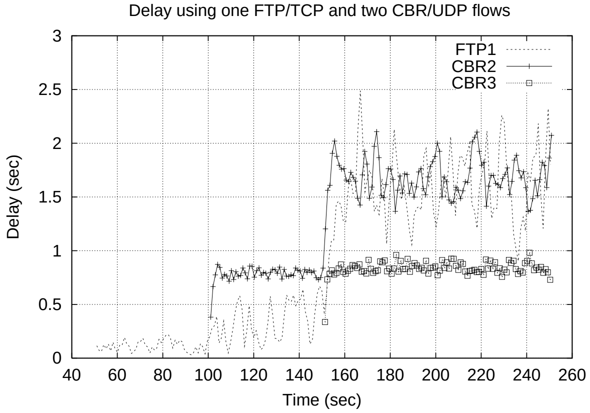


Figure 8.22: Mean Delay Using One TCP/FTP Flow and Two UDP/CBR Flows with the DIFS Differentiation Mechanism

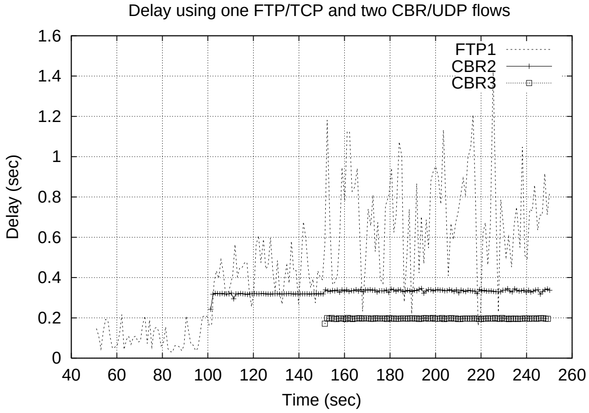


Figure 8.23: Mean Delay Using One TCP/FTP Flow and Two UDP/CBR Flows with the P3-DCF Mechanism

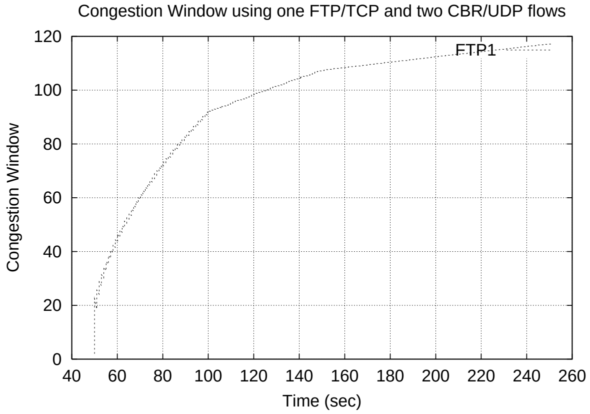


Figure 8.24: Congestion Window Using One TCP/FTP Flow and Two UDP/CBR Flows with the DIFS Differentiation Mechanism

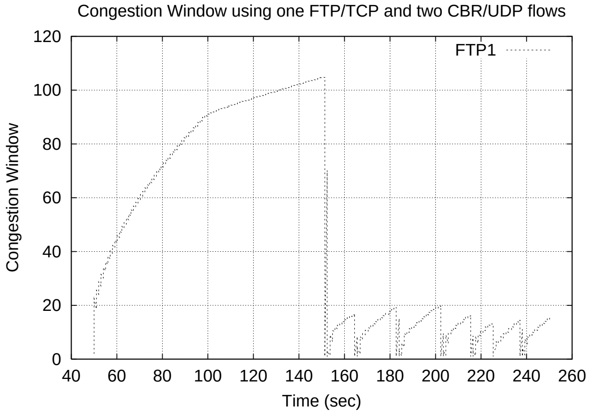


Figure 8.25: Congestion Window Using One TCP/FTP Flow and Two UDP/CBR Flows with the P3-DCF Mechanism

8.6.3.2 One flow UDP and Two flows TCP

In this scenario, three flows FTP1, FTP2 and CBR3 are generated such that FTP1 has lower priority than that of FTP2 which in turn has lower priority than that of the CBR3 flow. FTP1,

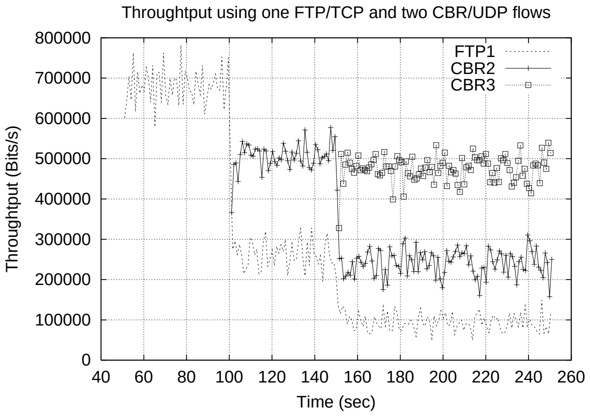


Figure 8.26: Throughput Using One TCP/FTP Flow and Two UDP/CBR Flows with DIFS Differentiation Mechanism

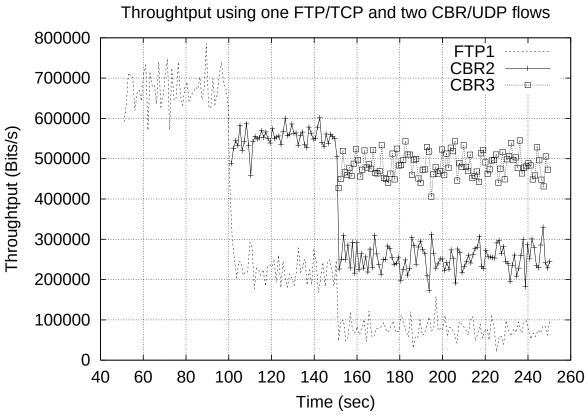


Figure 8.27: Throughput Using One TCP/FTP Flow and Two UDP/CBR Flows with the P3-DCF Mechanism

FTP2 and CBR3 start sending packets respectively at seconds 50, 100 and 150. The simulation parameters are given in Table 8.2 and Table 8.7.

	$\mathbf{DIFS_j^{min}/DIFS_j^{max}}$	$\mathbf{Temax_j}$
CBR3	$50\mu s/130\mu s$	$150ms$
FTP2	$130\mu s/210\mu s$	$750ms$
FTP1	$210\mu s/290\mu s$	$750ms$

Table 8.7: Simulation Parameters Using One UDP/CBR Flow and Two TCP/FTP Flows with the P3-DCF Mechanism

Figures 8.30 to 8.35 lead to the same conclusions as in the previous section. We can note that with P3-DCF, the UDP flow has priority over TCP flows and that the UDP flow's delay is limited and has small variation (figure 8.31). On the other hand, the congestion window of the lowest priority flow FTP2 enters frequently the slow start phase. As for FTP1, the slope of congestion window evolution decreases. However, this flow does not reenter the slow start phase (figure 8.33). As a consequence, a significant service differentiation is obtained with TCP flows with our mechanism (as in the scenario relative to the three TCP flows). Moreover, this differentiation is again clearly obtained between the UDP flow and the TCP flows.

8.7 Conclusion

In this chapter, we propose a novel mechanism called P3-DCF. This mechanism is based on the extension of the DCF wireless access method standardized by the IEEE 802.11 working group. The proposed extension gives a time-dependent priority to each packet, thus making the proposed mechanism behaves as an Earliest Deadline First discipline. Furthermore, the proposed mechanism is distributed and reduces the signaling load and the transmission overheads in comparison to a centralized mechanism.

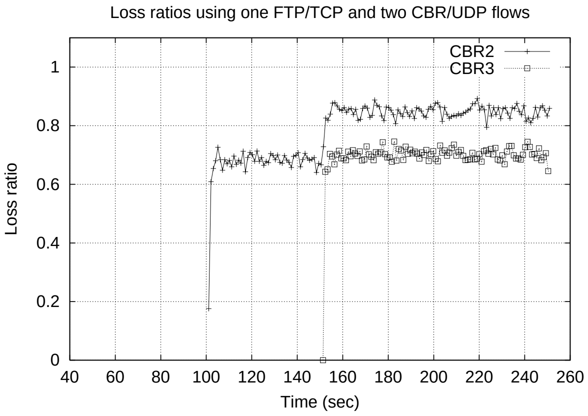


Figure 8.28: Loss Ratio Using One TCP/FTP Flow and Two UDP/CBR Flows with the DIFS Differentiation Mechanism

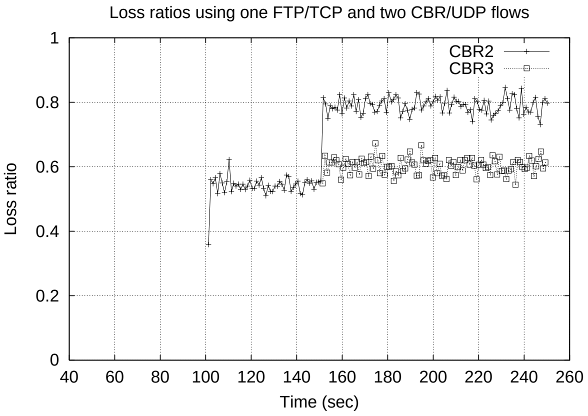


Figure 8.29: Loss Ratio Using One TCP/FTP Flow and Two UDP/CBR Flows with the P3-DCF Mechanism

The performance evaluation of P3-DCF is promising and demonstrates that it enhances the differentiation among flows and that it decreases the delays experienced by packets when compared to a simple DIFS differentiation. Furthermore, we show that, in contrary to previous DCF extensions proposed in the literature, our mechanism performs an effective service differentiation for UDP, TCP and mixed UDP/TCP traffic flows. In order to show the effectiveness of our mechanism in non saturated conditions, we plan to simulate and analyze the proposed mechanism using several scenarios and more accurate traffic models (voice, video, web, mail). Moreover, an admission control algorithm for UDP traffic flows is needed in order to keep the loss ratio below a predefined value.

This chapter has focused on the QoS provisioning in the MAC sub-layer in the IEEE 802.11 WLANs. In the next generation of mobile networks, access networks of different technologies will co-exist and will be connected to an IP core network. These access networks will be the WLAN networks as well as the GSM, GPRS, satellite and UMTS networks. In the previous and present chapters, we dealt with the quality of service in wireless IP and WLAN access networks. One interesting study is the Universal Mobile Telecommunication Systems (UMTS) quality of service study. This issue will be the subject of Chapter 9.

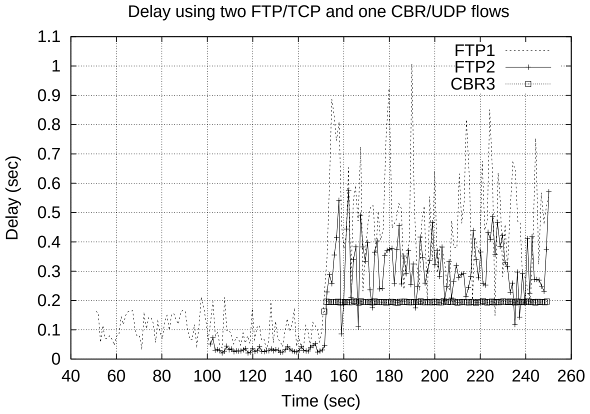


Figure 8.30: Mean Delay Using One UDP/CBR Flow and Two TCP/FTP Flows with the Proposed P3-DCF Mechanism

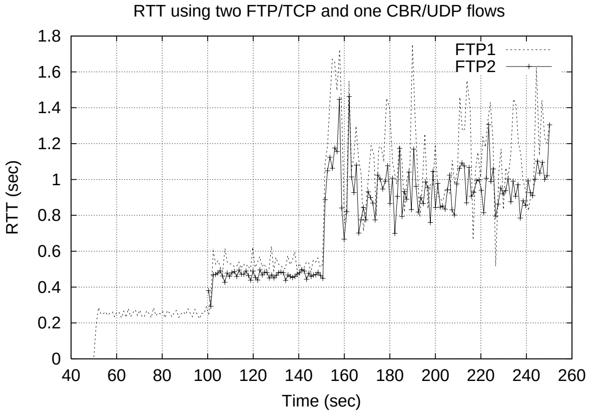


Figure 8.31: RTT Using One UDP/CBR Flow and Two TCP/FTP Flows with the Proposed P3-DCF Mechanism

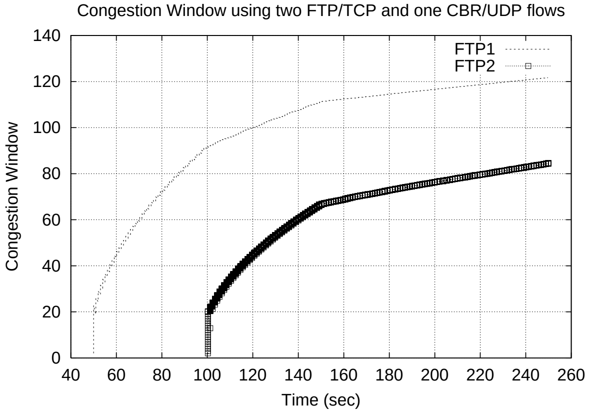


Figure 8.32: Congestion Window Using One UDP/CBR Flow and Two TCP/FTP Flows with the DIFS Differentiation Mechanism

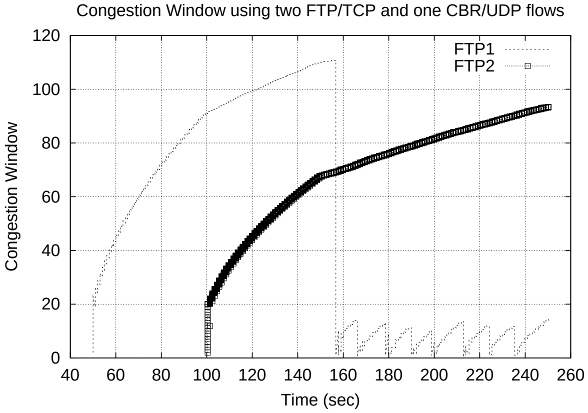


Figure 8.33: Congestion Window Using One UDP/CBR Flow and Two TCP/FTP Flows with the P3-DCF Mechanism

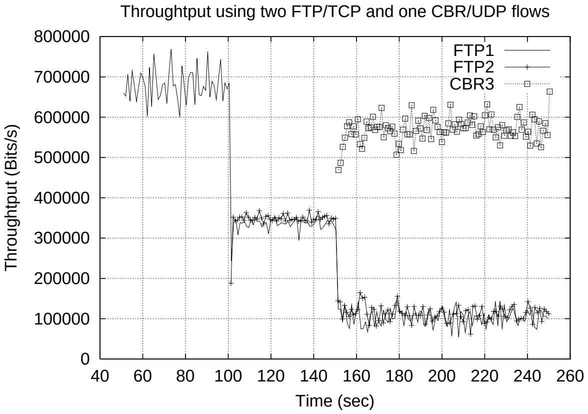


Figure 8.34: Throughput Using One UDP/CBR Flow and Two TCP/FTP Flows with the DIFS Differentiation Mechanism

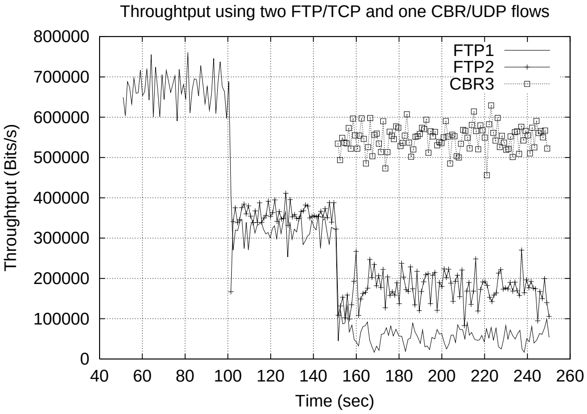


Figure 8.35: Throughput Using One UDP/CBR Flow and Two TCP/FTP Flows with the P3-DCF Mechanism

Chapter 9

A QoS Study of the MAC Layer in UMTS

9.1 Introduction

In the future generation of mobile networks, access networks of different technologies will co-exist and will be connected to an IP core network. Some of these access networks are GSM, GPRS, WLAN, satellite, wireless IP and UMTS networks. In the previous chapters, we have made QoS studies in WLAN and wireless IP networks. In the present chapter, we are concerned about the QoS study in the UMTS network.

Beyond the Release 99 of the UMTS networks based on the ATM technology [3rd], the mobile network will use IP. Within the UMTS access network, the MAC layer has been specified in order to provide QoS to mobile users. The MAC layer QoS will be based on the QoS provided by the underlying IP network.

The MAC layer has a significant impact on the user traffic. Thus, it affects the overall performance. This impact is more perceptible in the case of bursty data traffic such as the Web traffic.

In order to study the MAC layer, it is important to have a close look at the Radio Resource Control layer (RRC) which has a global view of the cells. RRC is responsible for the call admission control. Moreover, RRC can notice if there is a risk of congestion and acts proactively by reconfiguring the MAC layer in such a way that the MAC reduces the throughputs of the active users that it controls.

In the present chapter, we make a study of the MAC layer in the downlink. To this end, we start our study by defining the radio access network architecture in UMTS and by outlining the various tasks performed by the RRC and the MAC layer. Next, we present a model of a cell in UMTS. Then, a scheduling discipline will be detailed before presenting the simulation scenarios and results. The general conclusion and proposition of improvements are given at the end of the chapter.

9.2 UMTS Terrestrial Radio Access Network Architecture

The UMTS Terrestrial Radio Access Network (UTRAN) is similar to that of GSM. The architecture of UTRAN is illustrated in figure 9.1. The different elements of UTRAN are:

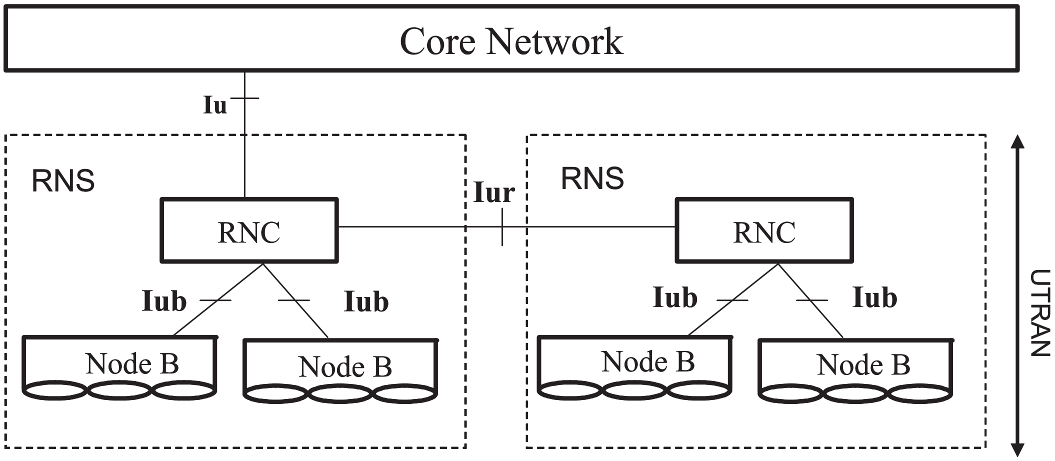


Figure 9.1: UTRAN Architecture

- The Radio Network Controller (RNC) is responsible for establishing/maintaining/releasing of the radio connection with the user equipment. RNC is also in charge of the call admission control and of the handover management.

- The Node B is similar to the Base Transceiver Station (BTS) in GSM. This node is in charge of transmitting the data and of the signaling over the radio interface.

- The Radio Network Subsystem (RNS) consisits of a RNC and one or more Node Bs. It is the access part of the UMTS network that permits the allocation and the release of the radio resources specific to a set of cells in order to establish the connections between the mobile and the UTRAN. Note that there is one RNC per RNS.

The different interfaces of UTRAN are:
Iu: it is the interface between a RNS and the core network.
Iub: it is the interface between the RNC and the Node B.
Iur: it is the interface between two RNCs.
These interfaces are IP-based in all IP UMTS networks.

Three layers can be distinguished in the UTRAN and the users equipment (UE) part (figure 9.2). These are the Physical Layer (L1), the Layer 2 which comprises the MAC and the Radio Link Control layer (RLC) layers, and the RRC layer. In our study, we are interested in the RRC and in the MAC layer.

9.2.1 RRC functions

The RRC layer handles the control plane signalling of Layer 3 between the UEs and UTRAN. The RRC performs the following functions [3rd02]:

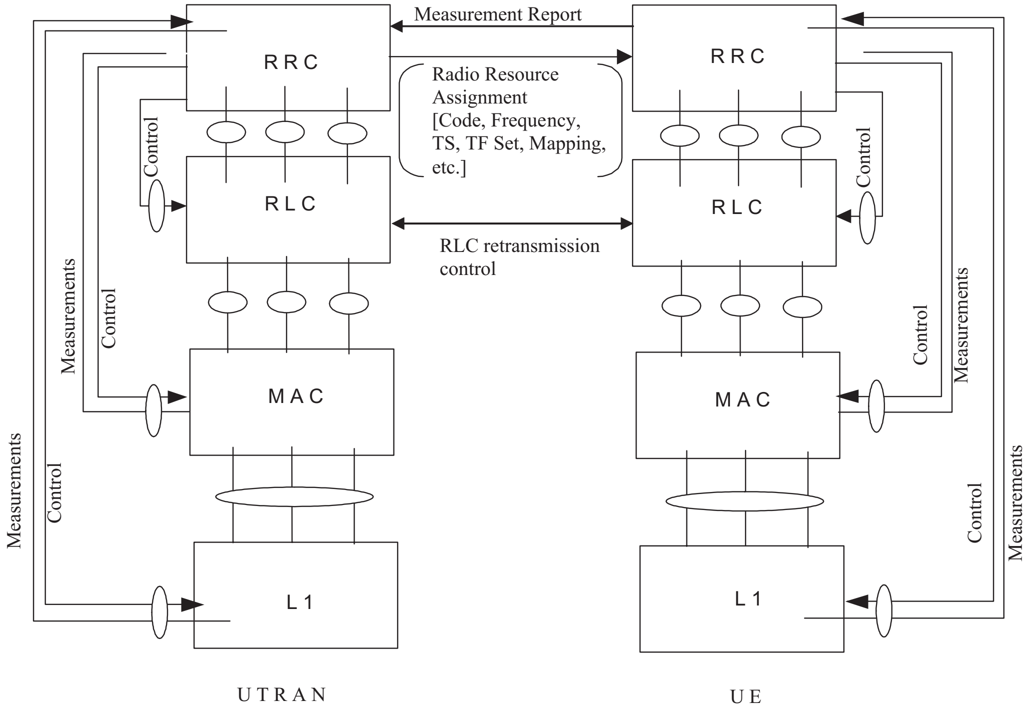


Figure 9.2: Interaction between the RRC layer and the other layers

- Broadcast of information provided by the non-access stratum (Core Network). The RRC layer performs system information broadcasting from the network to all the users equipments (UEs). The system information is normally repeated on a regular basis. The RRC layer performs the scheduling, segmentation and repetition. This function supports broadcast of higher layer (above RRC) information. This information may be cell specific or not.

- Broadcast of information related to the access stratum. The RRC layer performs system information broadcasting from the network to all UEs.

- Establishment, re-establishment, maintenance and release of an RRC connection between the UE and UTRAN. The establishment of an RRC connection is initiated by a request from higher layers at the UE side to establish the first Signalling Connection for the UE. The establishment of an RRC connection includes an optional cell re-selection, an admission control, and a layer 2 signalling link establishment. The release of an RRC connection can be initiated by a request from higher layers to release the last Signalling Connection for the UE or by the RRC layer itself in case of RRC connection failure. In case of connection loss, the UE requests re-establishment of the RRC connection. In case of RRC connection failure, RRC releases resources associated with the RRC connection.

- Assignment, reconfiguration and release of radio resources for the RRC connection. The RRC layer handles the assignment of radio resources (e.g. codes) needed for the RRC connection including needs from both the control and user plane. The RRC layer may reconfigure radio resources during an established RRC connection. This function includes coordination of the radio resource allocation between multiple radio bearers related to the same RRC connection. RRC controls the radio resources in the uplink and downlink. RRC signals to the UE to indicate resource allocations for purposes of handover to GSM or other radio systems.

- RRC connection mobility functions. The RRC layer performs evaluation, decision and execution related to RRC connection mobility during an established RRC connection, such as handover, preparation of handover to GSM or other systems, cell re-selection and cell/paging area update procedures, based on e.g. measurements done by the UE.

- Control of requested QoS. This function ensures that the QoS requested for the Radio Bearers can be met. This includes the allocation of a sufficient number of radio resources.

- UE measurement reporting and control of the reporting. The measurements performed by the UE are controlled by the RRC layer, in terms of what to measure, when to measure and how to report, including both UMTS air interface and other systems.

- and others...

The RRC protocol controls and signals the allocation of radio resources to the UE. The RRC allows MAC to arbitrate between users and Radio Bearers within the radio resource allocation. The RRC uses the measurements done by the lower layers to determine which radio resources are available.

In this subsection, we have detailed the main functions of RRC in order to point out the importance of the Radio Resource Control in the QoS provisioning. In our study, we shall take into account the interaction between the RRC and the MAC layer in order to study the QoS offered to the UMTS users.

9.2.2 MAC functions

This section provides an overview on services and functions provided by the MAC layer to the upper layers. The interested reader might want to have a look at [3rd00a].

- Data transfer. This service provides unacknowledged transfer of MAC Service Data Units (SDUs) between peer MAC entities. This service does not provide any data segmentation. Therefore, segmentation/reassembly function should be achieved by upper layers.

- Reallocation of radio resources and MAC parameters. This service performs on request of RRC execution of radio resource reallocation and change of MAC parameters, i.e. reconfiguration of MAC functions such as change of identity of UE, change of transport format (combination) sets, change of transport channel type.

- Reporting of measurements. Local measurements such as traffic volume and quality indication are reported to RRC.

- and others...

9.2.3 Classes of Service

In the 3GPP specification [3rd00b], four different QoS bearer classes were defined:

- *Conversational class* (e.g. telephony speech): it is intended to be used to carry real-time traffic. The fundamental characteristics for this class of service are to preserve time relation (variation) between information entities of the stream and to ensure a conversational pattern (stringent and low delay).
- *Streaming class* (e.g. real time video): it is intended to be used to carry real-time traffic. The fundamental characteristics for this class of service are to preserve time relation (variation) between information entities of the stream.
- *Interactive class* (e.g. web): the fundamental characteristics for this class of service are to ensure the request response pattern (delay) and to preserve the payload content.
- *Background class* (e.g. email): the fundamental characteristic for this class of service is that the destination is not expecting the data within a certain time. On the other hand, this class of service should preserve the payload content.

We limited our MAC QoS study to two classes of service: the first will bear the real time UMTS classes (conversational and streaming classes) and the second will bear the non real time classes (interactive and background classes). In our study, we analysed the voice and the Web traffic handled by the first and the second MAC classes.

Given a cell bandwidth equal to 732Kb/s, we try to evaluate the performance in terms of:

- Total packet delay (mean and maximum)
- Jitter for the voice traffic
- Radio bandwidth utilization ratio (mean)

9.3 Cell Model

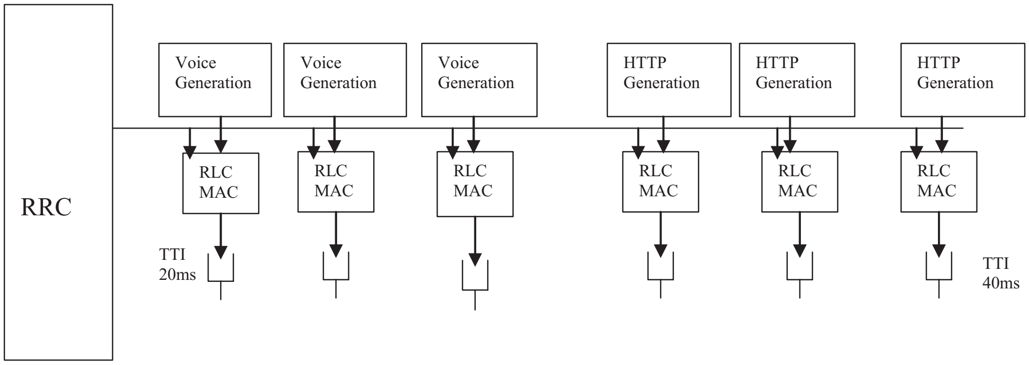


Figure 9.3: Model of a Cell in UMTS

The studied cell is depicted in figure 9.3. The cell comprises voice traffic generators as well as Web traffic (UDD 64 Kb/s) generators. The adopted traffic models are detailed in chapter 1.

The generated traffic is treated first by the RLC (figure 9.4). In the RLC layer, the generated packets are segmented in Transport Blocks (TB). The size of transport blocks depends on the expected radio performance as well on the desired bit rate and on the Transmission Time Interval (TTI).

The RLC segments the Web traffic to TBs of 80 bytes. Then, the RLC adds a header of two bytes to each. On the other hand, RLC is transparent for the voice traffic. Note that voice packets have fixed size and arrive every 20ms.

The MAC is in charge of collecting a number of RLC PDU. Note that the RLC PDU is also called the Transport Block. As MAC receives the number of TBs to be sent in one Time Transmission Interval from RRC, MAC forwards this information to RLC so that RLC prepares the required data.

The MAC has to fill the TTI without exceeding the allowed bit rate [Pro99]. In fact, the product of the number of transmitted TBs per TTI and the size of a TB must be less than or equal to the product of the negotiated bit rate and TTI .

On the other hand, the respect of time constraint is another problem. Thus, it is important to take into account the delay and the jitter in the scheduling algorithm. We shall see in the simulations results that respecting the constraint of not exceeding the allowed bit rate will induce an increase in the delay and a decrease in the bandwidth use.

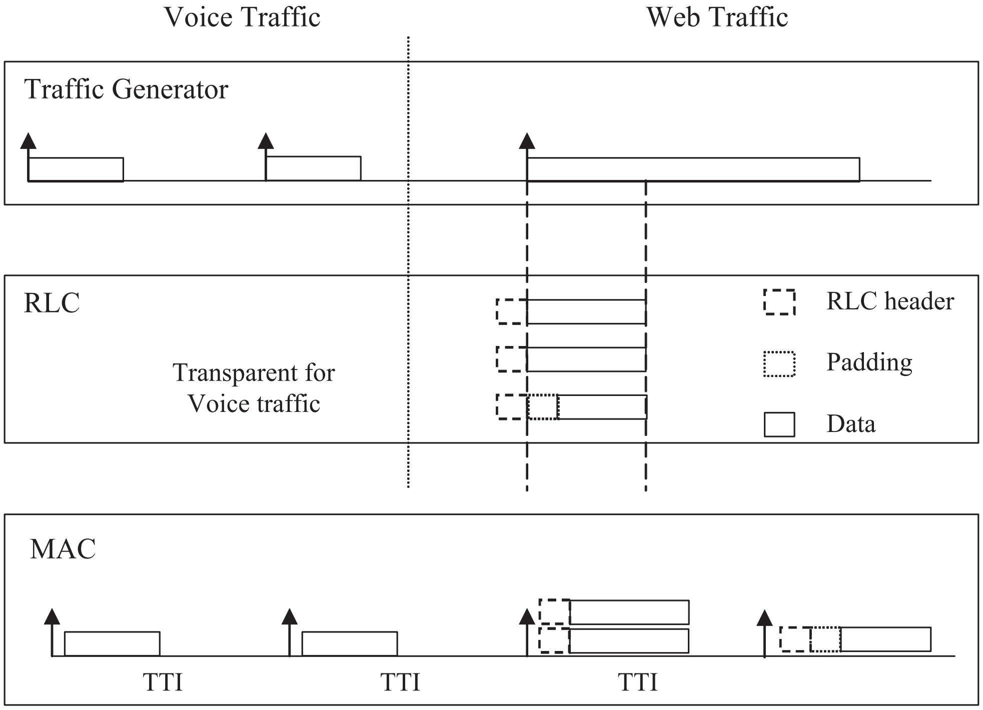


Figure 9.4: Processing of the Generated Traffic by the RLC and the MAC Layer

9.4 Scheduling Scheme

We suppose that the scheduler is located in the MAC layer. When RRC decides about the *Transport Format Combination Set (TFCS)* that can be used by the users, it informs the MAC layer about its decision. We represent TFCS as a set of resources available to the users. The resources can be the codes, the power needed to transmit the data, or the radio bandwidth. In our study, we consider that the resources are the radio bandwidth and more specifically the available transport blocks.

Let us define TFCS as $TFCS = (a_1, \ldots, a_n)$, where a_i represents the resources needed by the users. The MAC layer must decide about the a_i to allocate to the users. Every TTI, the MAC layer inspects the users buffers and the available resources and chooses one TFC to allocate to the users (figure 9.5).

The Transport Format Combination Set is what is given to MAC for control. However, the assignment of the Transport Format Combination Set is done by the Layer 3. When mapping data onto Layer 1, MAC chooses between the different Transport Format Combinations (TFC) given in the Transport Format Combination Set.

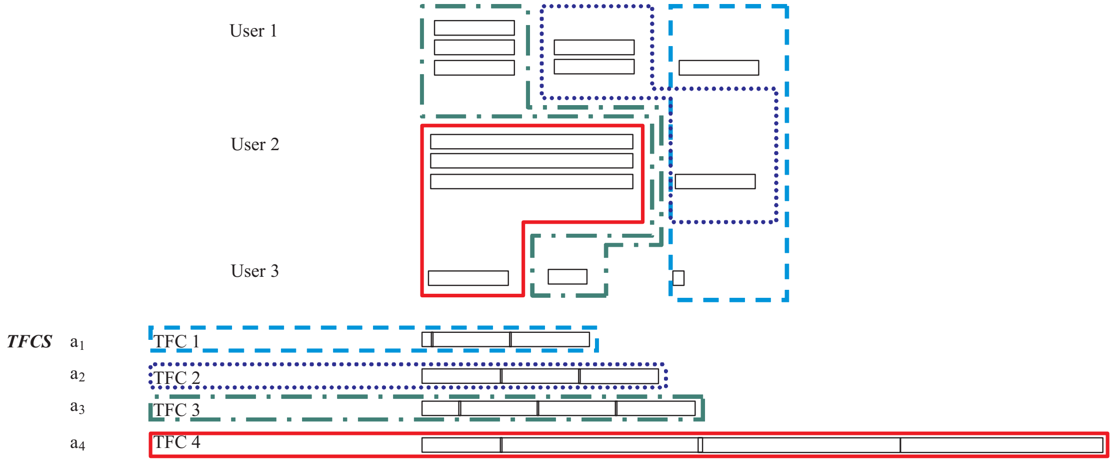


Figure 9.5: MAC Layer Scheduling

The selection of the Transport Format Combinations can be seen as a fast part of the radio resource control. The dedication of these fast parts of the radio resource control to MAC, close to L1, means that the flexible variable rate scheme provided by L1 can be fully utilised. These parts of the radio resource control should be distinguished from the slower parts, which are handled by the Layer 3 (RRC). Thereby the bit rate can be changed very fast, without any need for L3 signalling.

In the scenario that generates both voice and data traffic, the scheduler is activated each 20ms. In the case of only having data users and in the case of a TTI equal to 40ms, the scheduler is activated every 40ms.

The cell bandwidth is equal to 732Kb/s. Thus, it is equal to 1830 bytes available each 20ms.

These bytes must be shared between the voice and the data users.

For voice traffic, it is easier to evaluate the resource needed as voice packets are transmitted every 20 ms. For data traffic, like a web session, it is much more difficult as statistical effect has to be taken into account due to large silence period corresponding to the reading analysis of a web page.

Before detailing the adopted scheduling discipline, let us define the following parameters:

- N_r is the number of available TBs that can be transmitted.
- N is the total number of TBs waiting in the queues.
- N_{TBmax} is the maximum number of TBs that can be transmitted per Web user per TTI. Typically, it is equal to 4 for a WWW traffic of UDD 64Kbit/s.
- *nb_http_user* is the number of Web users.

Since, the voice traffic has the priority over the data traffic, the scheduler sends every 20ms all the voice packets waiting in the buffers. Note that the scheduler sends one voice TB per user per TTI. The unused bandwidth is then shared between the data users in the following set of rules:

- If N is less than N_r, the TBs waiting in the buffers will be totaly transmitted if the number of transmitted TBs per user does not exceed N_{TBmax}.
- if N is greater than N_r, then two cases are to be distinguished:
  1. If N_r is less than *nb_http_user*, the Earliest Deadline First (EDF) is then applied: the priority is given to the TB with the deadline that expires the first. This TB is selected for transmission.
  2. Otherwise, each user can transmit N_r/*nb_http_user* TBs at the most. If one (or more) user does not have data to transmit, then the unused bandwidth will be shared between other users according to the EDF discipline.

9.5 Simulation and Proposed Scenarios

The simulation of the cell depicted in figure 9.3 was carried out using OMNeT++. We intended to show the performance of the MAC layer in an UMTS cell, using the proposed scheduling scheme. The performance measures were retrieved for the four following cases:

1. 100% of the cell users are voice users.
2. 100% of the cell users are data users.
3. 80% of the cell users are voice users and 20% of the cell users are data users.
4. 80% of the cell users are data users and 20% of the cell users are voice users.

Note that the above-mentionned percentages represent the percentages of the users number and not the percentages in term of the bandwidth occupation.

9.5.1 Case 1: 100% Voice Users

We suppose that there is a synchronisation between the MAC layer and the voice packets generation: voice packets are generated every TTI.

Figures 9.6 and 9.7 show that the voice delay and voice jitter are negligible until the number of users reaches the value of 100. The delay and the jitter start to grow when the users number becomes greater than 100 and become unacceptable when there are 110 active users in the cell. When the number of users is 108, the maximum delay reaches 20ms. For 110 users, the retrieved maximum delay becomes unacceptable. As a result, we can deduce the following conclusion: the cell can accept at the most 107 voice users if the maximum delay should be kept below 20ms.

Note that in the cases 3 and 4 where there are simultaneously voice and data users, the voice delay is not affected by the data traffic, since the voice traffic has the priority over the data traffic. *Consequently, the measured delays in the next sections concern the data packets.*

Figure 9.8 shows the mean and the maximum percentage of the used bandwidth. The depicted curves indicate that the percentage of the used bandwidth grows with the load increase.

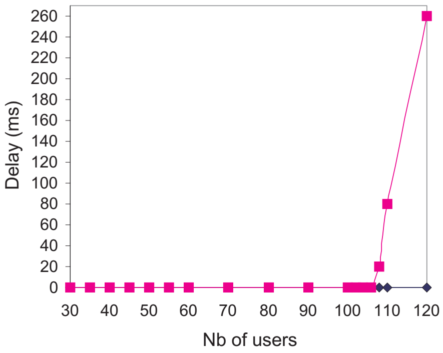


Figure 9.6: Voice Delay: 100% Voice Users, mean delay (resp. maximum delay) is shown by the curve with the diamond (resp. square points)

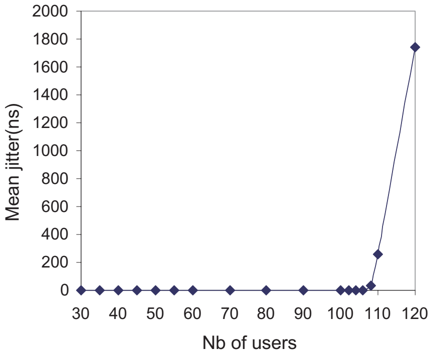


Figure 9.7: Voice Jitter: 100% Voice Users

9.5.2 Case 2: 100% Data Users

For data traffic, the experienced delays grow with the increase of the number users. First simulations were carried out with a maximum size of the HTTP packets equal to 64 kbytes as stated in the ETSI specification (see chapter 1). We have then obtained very large delays that reach 12s for 80 users (figure 9.9). This result was expected.

In fact, the maximum size of the data packets being equal to 64 kbytes, the number of TBs segmented from a data packet that has the maximum size will be 800. In the case when the MAC layer permits to have a throughput equal to 64kb/s, the minimum time necessary for the transmission of a data packet of 64 kbytes is 8s. It is noteworthy that this is an optimistic case

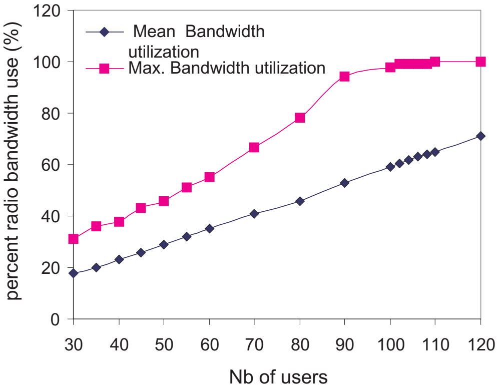


Figure 9.8: Percentage of the bandwidth use: 100% Voice Users

because for high load, the MAC layer sends less than 4 TBs per TTI. Consequently, the last TB to be transmitted must wait for a longer time in the buffers.

In the terrestrial systems the transport layer segments the HTTP packets in small segments. Consequently, the case of having a maximum size equal to 64 kbytes is not a realistic case. Thus, we can draw the following conclusion: *the values of the traffic model parameters that were suggested in the ETSI specification do not lead to the required performance.*

Therefore, we have reduced the maximum size of the data packet. We have adopted a size of 10 kbytes (FDDI networks) and then a size of 1502 bytes (local area networks). With these values, the delays must considerably decrease. In fact, we must expect a maximum delay equal to 0.125s for a size of 10kbytes. For a size of 1502 bytes, this delay will be equal to 0.19s. These delays will be greater than these values in case of high load. Our expectations were confirmed by the achieved simulations (figures 9.10 and 9.11).

The fact that we were obliged to send at the most N_{TBmax} data TBs per TTI penalizes the bandwidth use. This maximum number of TBs was a constraint imposed by the Minicell project that supports this work. This was done in order not to exceed the data UDD as explained in section 9.3. This leads to small value of the radio bandwidth use and to relatively high values of delay (figure 9.12).

9.5.3 Case 3: 80% Voice Users, 20% Data Users

In this case, for low loads the delays fluctuate slightly (figures 9.13 and 9.14). In fact, when the number of data users is small, the total throughput of the data users is very small and the delays have small variation. When the number of users reach 120 in the case of the maximum HTTP packet size equal to 10 kbytes, the delays start growing. The percentage of the used bandwidth is greater than in the case of a system that contains 100% data users (figure 9.15). This is due to the great number of voice users that have the priority to access the medium.

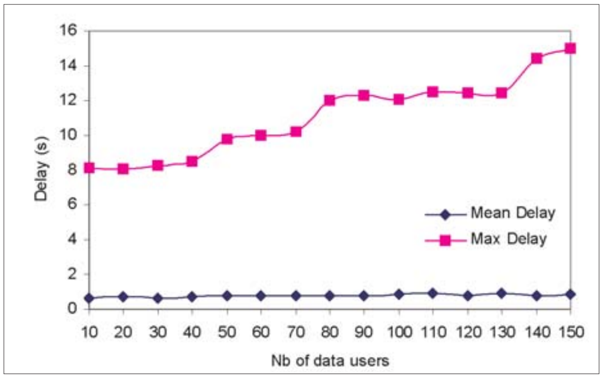


Figure 9.9: Mean and Maximum Data Delay: 100% Data Users, Maximum size of HTTP Packet 64 kbytes

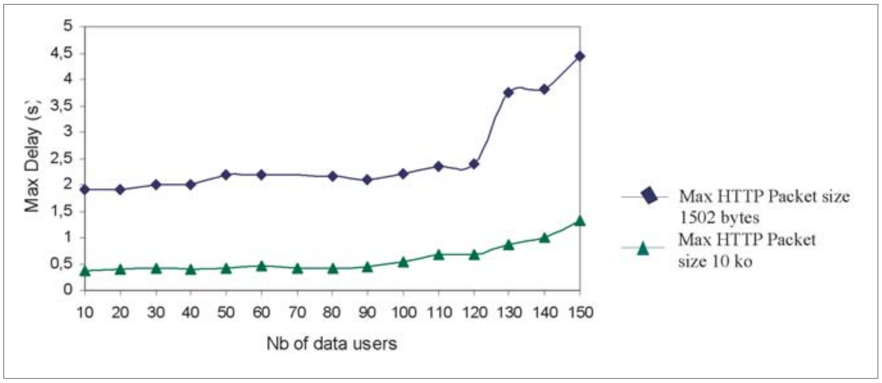


Figure 9.10: Maximum Data Delay: 100% Data Users, Maximum size of HTTP Packet 10 kbytes and 1502 bytes

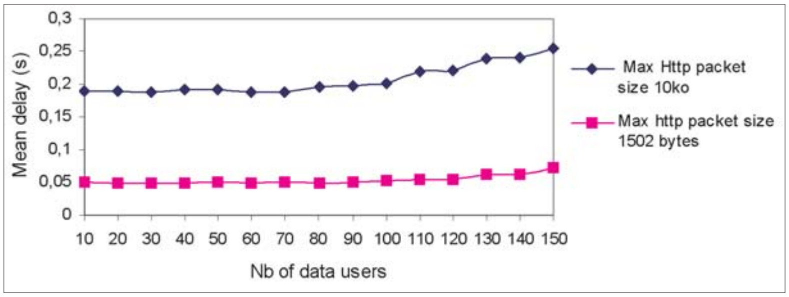


Figure 9.11: Mean Data Delay: 100% Data Users, Maximum size of HTTP Packet 10 kbytes and 1502 bytes

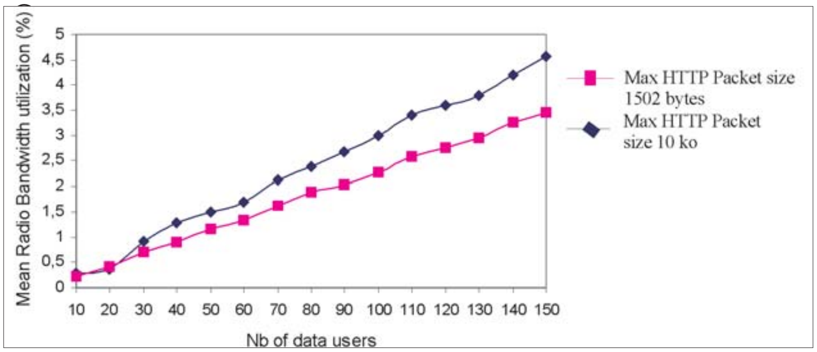


Figure 9.12: Mean Utilization Bandwidth: 100% Data Users, Maximum size of HTTP Packet 10 kbytes and 1502 bytes

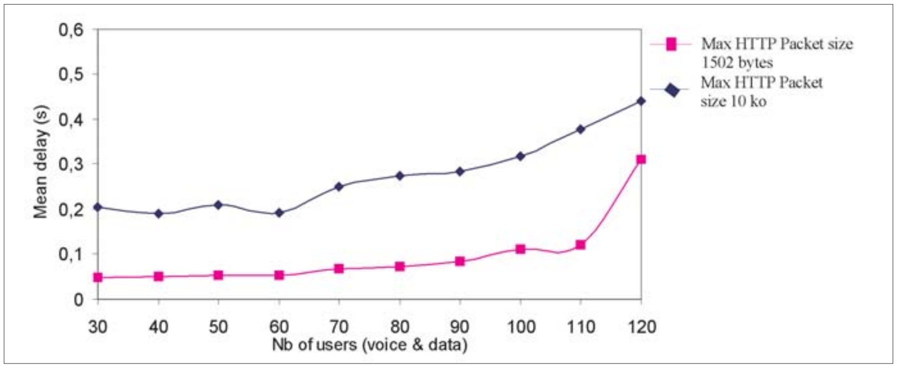


Figure 9.13: Mean Delay: 80% Voice Users, 20% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

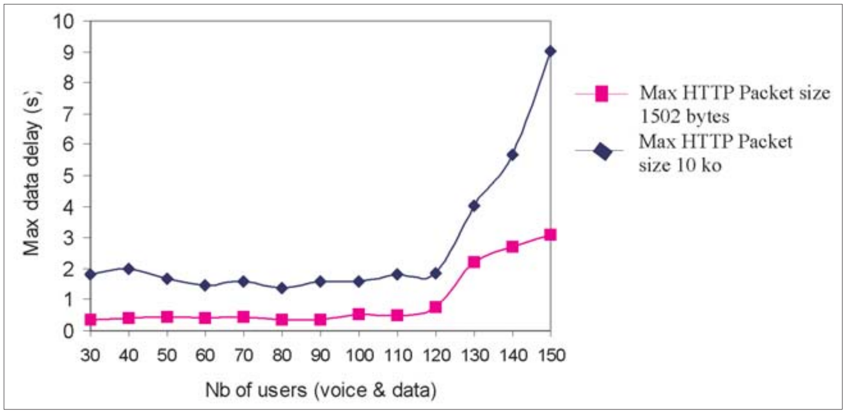


Figure 9.14: Maximum Delay: 80% Voice Users, 20% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

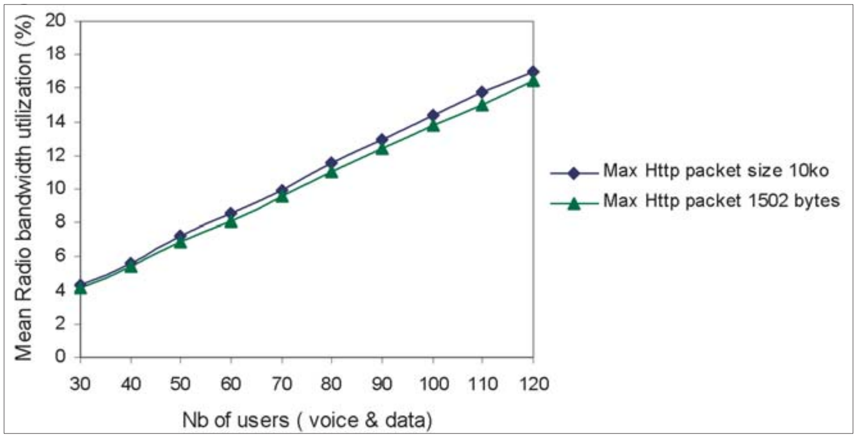


Figure 9.15: Mean Bandwidth: 80% Voice Users, 20% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

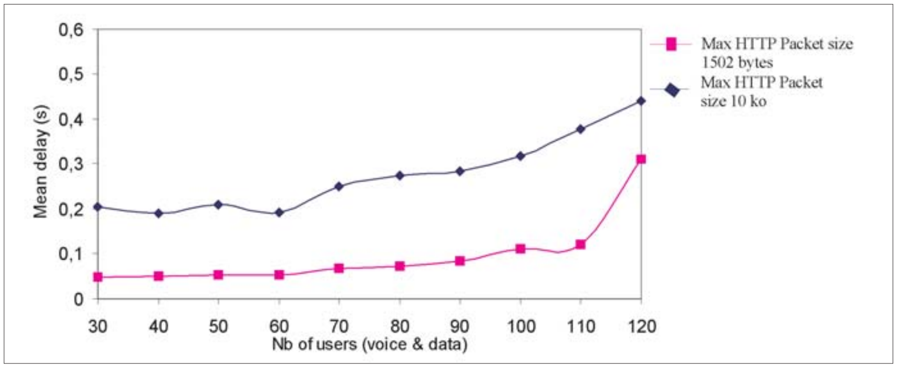


Figure 9.16: Mean Delay: 20% Voice Users, 80% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

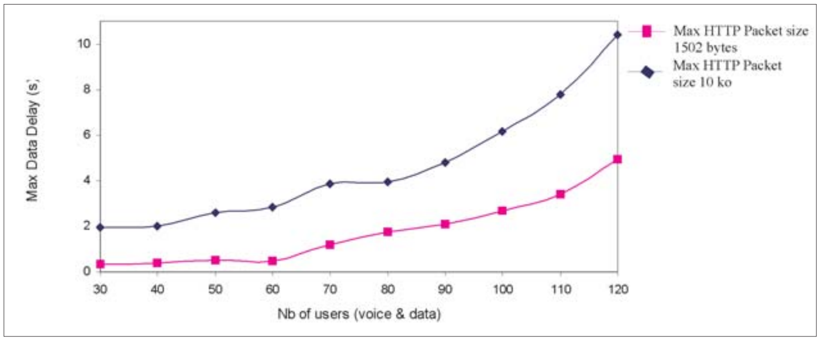


Figure 9.17: Maximum Delay: 20% Voice Users, 80% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

9.5.4 Case 4: 80% Data Users, 20% Voice Users

In this case, simulations were carried out in order to show the delay and the percentage of the used bandwidth. As we can see, the mean and the maximum delay grow with the load increase (figures 9.16, 9.17). We could have obtained smaller delays and greater bandwidth utilization (figure 9.18) if we have increased the number of maximum data TBs permitted to be transmitted.

9.6 Performance Improvement

In order to improve the quality of service of the system and to reduce the experienced delays, we have suggested two propositions:

1. The data packets with expired delays are dropped from the waiting queues: the packets with delays greater than 0.5s are dropped. Thus, these packets will not penalize other packets waiting in the buffers. In this case, we compute the loss probability.

2. Another method to reduce the data experienced delays is to reduce the TTI value while

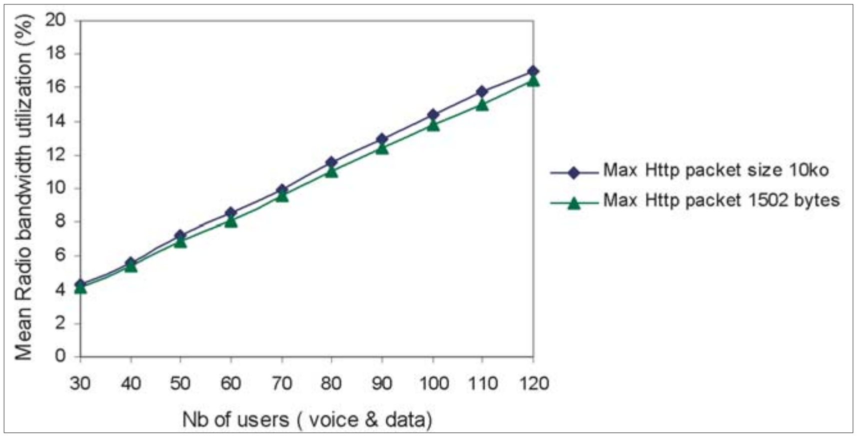


Figure 9.18: Mean Bandwidth: 20% Voice Users, 80% Data Users, Maximum size of HTTP Packet 10 kbytes, 1502 bytes

reducing the TB size. For an UDD equal to 64 kb/s, a TTI equal to 40 ms and a maximum number of 4 TBS to transmit per TTI, the size of a TB is equal to 80 bytes while respecting the following formula:

$$UDD = \frac{N_{TB} S_{TB}}{TTI} \tag{9.1}$$

where N_{TB} is the number of TBs to be transmitted per TTI and S_{TB} the TB size. In order to preserve an UDD equal to 64 kb/s, we must reduce TTI while reducing the TB size. In our simulations, we have taken a TTI equal to 20ms and a size of 40 bytes.

Simulations were carried out for both suggestions, with a maximum data packet size equal to 1502 bytes. We considered two cases:

1. 100% of the cell users are data users.

2. 80% of the cell users are voice users and 20% of the cell users are data users.

9.6.1 Case1: 100% Data

With the first proposition, the performance obtained in terms of delay are improved (figures 9.19 and 9.20). This is done at the expense of a sligth increase of the loss probability (figure 9.21). Note that the loss probability is relatively low.

With the second proposition, we do not notice a significant difference in the performance obtained with the two values of TTI (20 and 40ms) for low loads. In fact, with small loads, the radio bandwidth is not sufficiently loaded and thus a frequent resource allocation will not necessarily improve the performance. The difference between the two curves (TTI= 20ms, TTI=40ms) is noticed at high loads. In fact, for a TTI equal to 20ms, the resource allocation will be more dynamic than a resource allocation done every 40ms (figures 9.19 and 9.20).

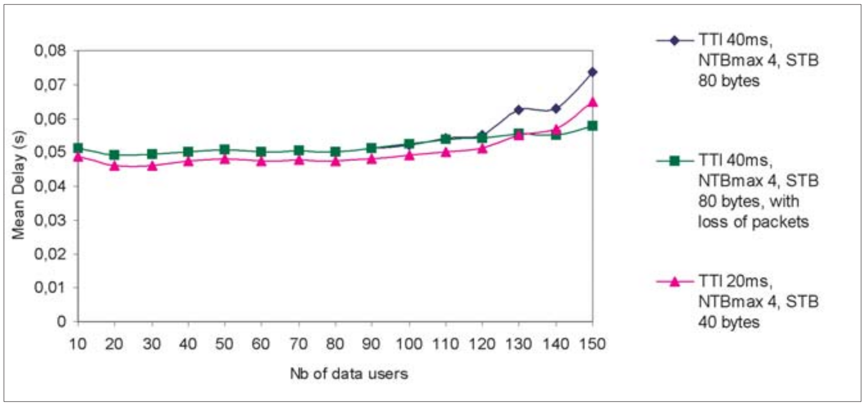


Figure 9.19: Mean Delay: 100% Data Users, Maximum size of HTTP Packet 1502 bytes (Perfomance Improvement)

9.6.2 Case2: 80% Voice 20% Data

Figures 9.22 and 9.23 show the delays obtained with the two strategies. The depicted curves show that for low loads, the performance obtained does not change from one scenario to another. However, when the number of users increases, the delays begin to increase. With 120 data users, the delay increase with a TTI equal to 40ms and a TB size equal to 80 bytes. With both propositions, the delays are rather stable with the load increase (figure 9.22) .

We again notice that the performance improvement obtained with the first proposition is achieved without inducing a perceptible loss probability (figure 9.25).

9.7 Conclusion

In this chapter, we studied the QoS issues provided by the UMTS MAC layer to mobile users. This study allows us to characterize the traffic after being treated by the MAC and thus this study characterizes the aggregated users' flows at the edge routers. This is an important task in the perspective of the performance evaluation of the core network.

We have analyzed the impact of the ETSI Web traffic model on the MAC performance. We showed that the value for the maximum allowed data packet size is inappropriate in order to guarantee the required QoS (data delay). We proposed to adopt a smaller maximum size of data packets equal to 1502 bytes. We showed that this maximum size is appropriate for the UMTS.

In order to further improve the data performance, we have proposed two strategies. The first strategy consists to drop the packet that present an expired delay. With this strategy, the data delays are reduced with a slight increase in the loss probability. As for the second strategy, we considered a TTI of 20ms and a TB size of 40 bytes to improve the data performance. This improvement is achieved because a dynamic resource allocation on the MAC layer leads to an efficient bandwidth allocation.

In our study, we obtained small bandwidth utilization in the case of a high percentage of data users. This is mainly due to the limitation of the maximum number of TBs that are permitted to be transmitted. We conclude that this constraint should be more flexible in order to improve

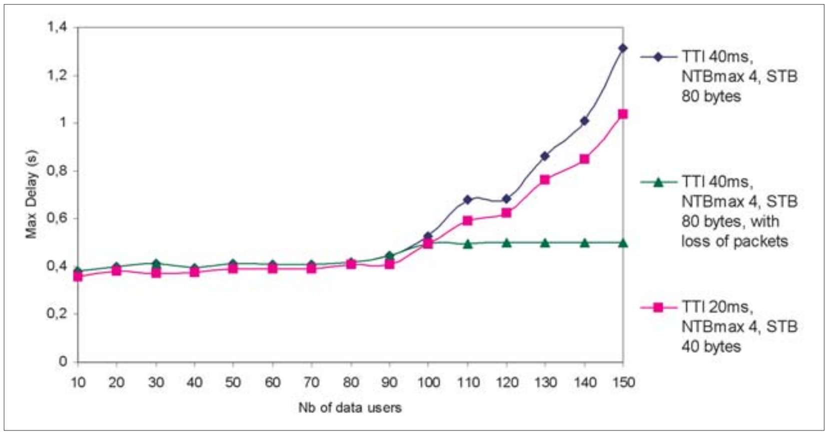


Figure 9.20: Maximum Delay: 100% Data Users, Maximum size of HTTP Packet 1502 bytes (Perfomance Improvement)

the bandwidth utilization.

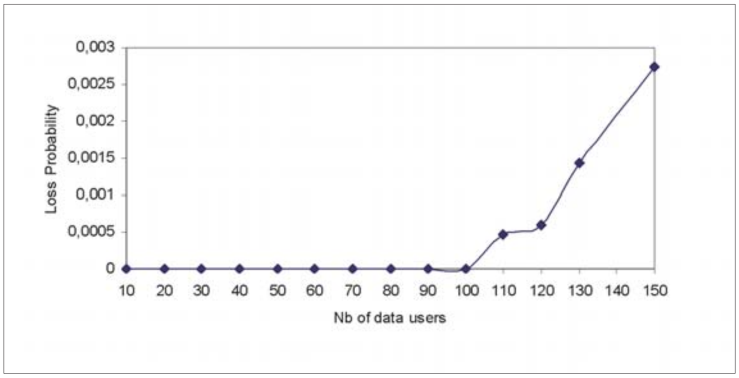


Figure 9.21: Loss Probability: 100% Data Users, Maximum size of HTTP Packet 1502 bytes, TTI 40ms, NTBmax 4, STB 80 bytes with loss (Perfomance Improvement)

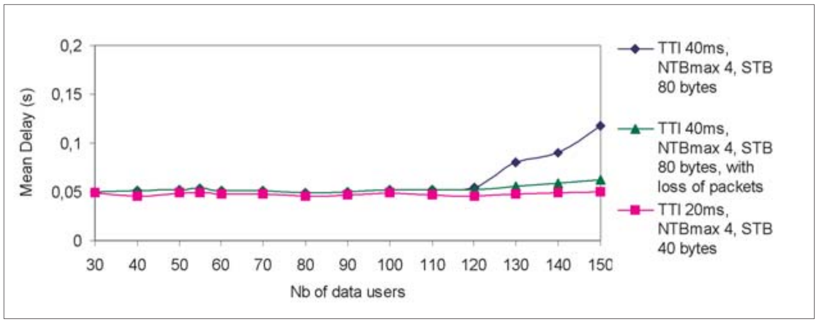


Figure 9.22: Mean Delay: 80% Voice Users, 20%Data Users, Maximum size of HTTP Packet 1502 bytes (Perfomance Improvement)

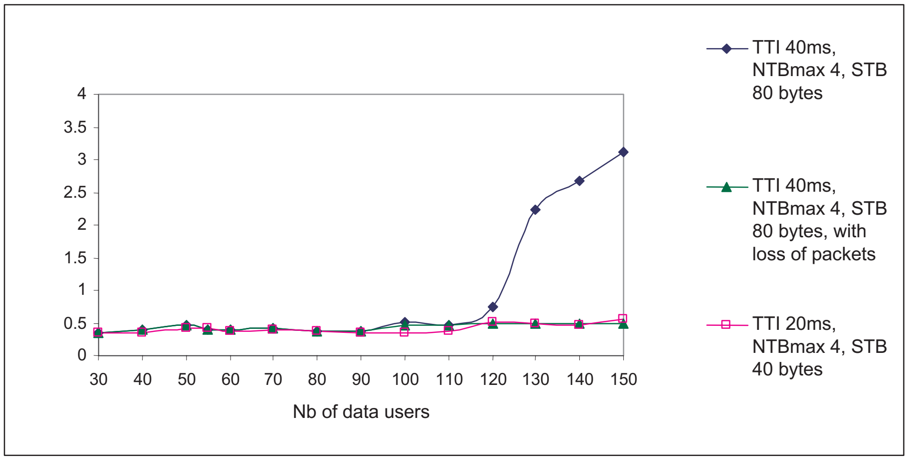


Figure 9.23: Maximum Delay: 80% Voice Users, 20%Data Users, Maximum size of HTTP Packet 1502 bytes

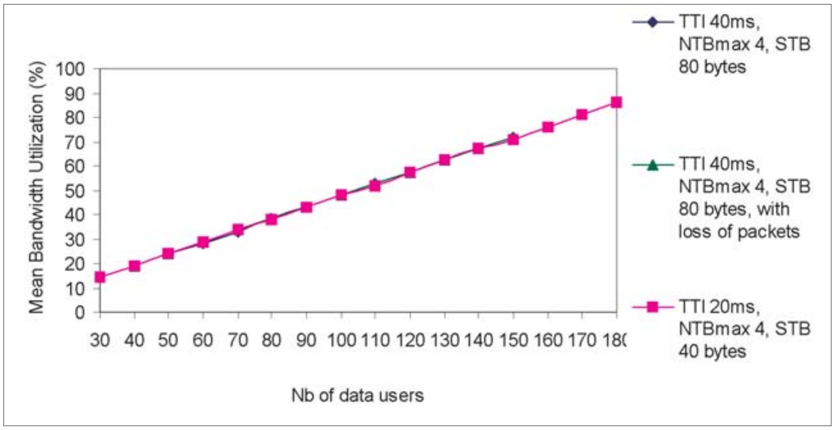


Figure 9.24: Mean Bandwidth: 80% Voice Users, 20%Data Users, Maximum size of HTTP Packet 1502 bytes (Perfomance Improvement)

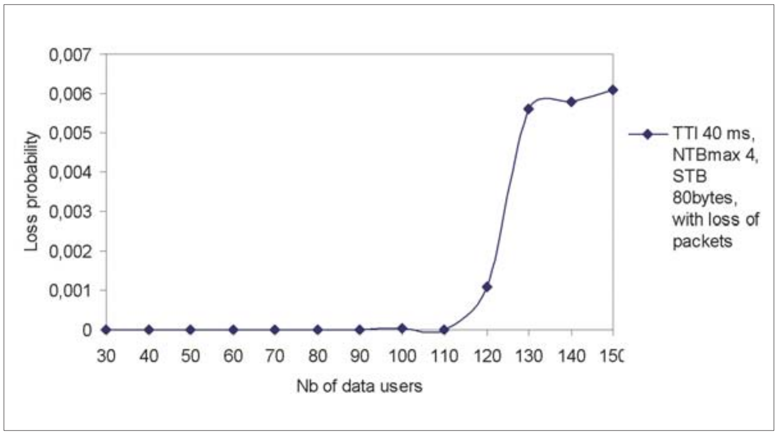


Figure 9.25: Loss Probability: 80% Voice Users, 20%Data Users, Maximum size of HTTP Packet 1502 bytes (Perfomance Improvement)

Chapter 10

General Summary and Conclusions

Provision of various real time multimedia services to mobile users is the main objective of current and next generation wireless networks. Major challenges in wireless mobile networks include support of fast handover and the provision of Quality of Service.

The main contribution of this thesis lies in the area of mobility handling and resource allocation procedures in wireless mobile multiservice networks. We aimed at offering different services to the mobile users while satisfying their QoS requirements and efficiently handling their mobility across different networks.

To this end, we started our work by proposing a threshold CAC with three priority levels supporting four classes of service. To maintain the handover priority, we assumed that the proposed scheme operates under a trunk reservation policy and a queuing strategy. Two different queuing disciplines have been introduced to further enhance the scheme characteristics: the non-preemptive priority Head of the Line (HOL) and the Queue Length Threshold (QLT). In addition to the analytical model developed in order to evaluate the performance of the proposed scheme, we have carried out simulations that validate the analytical model and further generalize the adopted assumptions. We found that with QLT, there is a significant improvement of data performance without inducing a perceptible degradation of the voice QoS. We also showed that QLT performs better when the number of guard channels C_h is smaller and when the number of channels reserved for new voice calls and for handovers C_v is higher.

In end-to-end QoS frameworks for multimedia wireless mobile networks, the high level of fluctuation in the availability of network resources is the major issue to be addressed. We believe that the wireless medium requires a fundamental change in the expectations we have from the service and the level of quality of service provided by the network. Multimedia applications need to be adaptive, to renegotiate the service request and to deal with changing conditions. On the other hand, end systems must be network aware as they must take the network status into account and must be able to adapt the multimedia streams accordingly. This adaptive approach implies that the network and the application are responsible for providing QoS to mobile users.

Thus, we have focused on developing a dynamic adaptive architecture DYNAA that provides this approach. The main feature of DYNAA is the capability to establish a collaboration between the application and the network. This collaboration handles the user's mobility and the high variability in network conditions while offering the best possible service to the user. Hence, the end-to-end QoS provisioning is a responsibility shared between the network and the application.

Within DYNAA, we have developed an admission control based on adaptive bandwidth algorithms that strive to provide QoS to two classes of service. The simulations show that by dynamically adjusting the allocated bandwidth while taking into account the current network conditions, the proposed DYNAA can be dynamic and consequently achieve better QoS. The extended version of DYNAA, namely *DYNAA_ Wait* helps in further enhancing the system performance at the expense of a slight degradation which is controlled and limited. The interesting results led us to integrate DYNAA in Cellular IP networks.

Because most probably next generation wireless networks will be IP-based and are expected to inter-work with the Internet backbone seamlessly, we proposed to study the "Cellular IPv6" protocol. Two major drawbacks of this protocol are the lack of support of an optimized uplink routing mechanism and provision of quality of service.

In a first step, we proposed to enhance the uplink routing mechanism. The proposed mechanism optimizes the routing of intra-network traffic, while minimizing the data loss during handover. Our mechanism minimizes the data packets delay, the signaling load, the data load and the handover establishment delay. These better results are obtained at the expense of some complexity added to the Cellular IP nodes. One must make the trade-off between better performance and complexity.

In a second step, we oriented our efforts towards the design of an efficient CAC that offers QoS to different classes of service in Cellular IPv6 networks. Thus, we integrated the dynamic adaptive architecture into a Cellular IPv6 network. Afterwards, we compared different types of CAC: centralized in the network gateway, distributed in the cells and hybrid referred to as centralized/distributed CAC scheme.

We showed the performance improvements achieved with the centralized/distributed CAC scheme, in terms of forced termination probability, blocking probability of new calls belonging to the Hard Adaptive class and average waiting delays. On the other hand, the distributed scheme has shown good performance as it reduces the load and the processing in the gateway. The distributed scheme can be efficient and flexible in the case of highly bursty traffic. Thus, there is a trade-off between the centralized/distributed and the distributed scheme.

On the other hand, we have been interested in providing an end-to-end QoS using Cellular IPv6 in the micro-mobility domains. Thus, we proposed an end-to-end QoS approach using IntServ in the Cellular IPv6 networks and DiffServ in the core network. The problems that exist when using DiffServ and IntServ in mobile environments were discussed. We then developed our proposed approach detailing the RSVP signaling during intra-network and inter-network handovers.

Since a significant interest for wireless local area networks is observed at the moment, we extended our work to the QoS issues in WLAN 802.11. Hence, we enhanced the MAC protocol defined within the IEEE 802.11 and introduced a novel mechanism that performs service differentiation. Our mechanism called, *P3-DCF*, uses the DCF as a fundamental access for prioritized service in IEEE 802.11 WLAN. An enhanced function, Per-Packet Priority (P3), integrated to the DCF mechanism establishes not only a per-flow differentiation but manages to schedule the packets with an Earliest Deadline First discipline as well.

The simulations carried out show that the proposed MAC mechanism satisfies the maximum tolerable latency of the real-time traffic and performs efficient flow differentiation. Moreover, the performance improvement is achieved without affecting the useful throughput.

Finally, we studied the MAC layer in UMTS. This study allows us to characterize the aggregated users' flows at the edge routers. This is an important task in the perspective of the performance evaluation of the core network.

We have analyzed the impact of the ETSI Web traffic model on the MAC performance. We showed that the value for the maximum allowed data packet size is inappropriate in order to guarantee the required QoS (data delay). We proposed to adopt a smaller maximum size of data packets equal to 1502 bytes. We showed that this maximum size is appropriate for the UMTS.

In order to further improve the data performance, we have proposed two strategies. The first strategy consists to drop the packet that present an expired delay. With this strategy, the data delays are reduced with a slight increase in the loss probability. As for the second strategy, we considered a TTI of 20ms and a TB size of 40 bytes to improve the data performance. This improvement is achieved because a dynamic resource allocation on the MAC layer leads to an efficient bandwidth allocation.

Perspectives

Within our study of the DYNAA architecture, we have limited our working field to the design of a call admission control based on adaptive bandwidth algorithms. Additional work can be pursued to permit the user's application to renegotiate and even to refuse the network decision. This signaling aspect between the network adaptive layer and the application layer is very important to study since it carries information between the user and the network.

On the other hand, traffic models of higher burstyness than the models adopted in the proposed architecture should be studied. The monitoring time interval ΔT should be adapted to the studied traffic models. A small time interval leads to a more precise application adaptation, but also increases the signaling load. Thus, there is a trade-off between the application adaptation, the frequency of the measurements and the signaling load. The fine tuning of the measurement time interval is an important task to be achieved by the network designer.

On the other hand, an interesting issue that is worthy to investigate is the the mobility handling for different users with different speeds within the Cellular IPv6 networks. In order to cope with terminals having different speeds and in order to achieve high user capacity, the "multilayer" architecture can be proposed for efficient management of radio resource. The proposed DYNAA has to cope with such an architecture. Handover issues with speed-sensitive algorithms that allow transfer of mobile stations between layers of the multilayer network are interesting to investigate.

The mechanisms and algorithms, developed all along this thesis, have been mainly targeting mobility and resource allocation within networks having the same technology. These proposed mechanisms and algorithms can be applied for second, third and fourth generation networks. Their applicability to networks having different technologies can also be envisaged, if complemented with the necessary handover functionalities. Thus, it is worthwhile to study the vertical handover in networks having different technologies.

Finally, the mobility handling and the resource allocation procedures in wireless mobile multiservice networks constitute a very vast research area. We hope that our studies will help the work in this domain. The modest contribution achieved throughout this thesis does not reduce the challenges that still remain to be overtaken for the next wireless generation networks.

Bibliography

[3rd99a] 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects. *Services and Service Capabilities*, Mars 1999.

[3rd99b] 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects. *Mandatory Speech Codec Speech Proceeing Functions; AMR Speech Codec; General Description*, August 1999.

[3rd00a] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network. *MAC Protocol Specification*, Mars 2000.

[3rd00b] 3rd Generation Partnership Project; Technical Specification Group Services and System Apects. *QoS Concept and Architecture*, Mars 2000.

[3rd02] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network. *Radio Interface Protocol Architecture*, September 2002.

[3rd] 3rd Generation Partnership Project; 3GPP Specifications. *http://www.3gpp.org/specs/numbering.htm.*

[AC01] I. Aad and C. Castelluccia. Differentiation mechanisms for ieee 802.11. In *IEEE INFOCOM*, volume 1, pages 209 –218, April 2001.

[AC03] I. Aad and C. Castelluccia. Priorities in wlans. *Computer Networks*, 2003.

[BBM01] G. Bianchi and N. Blefari-Melazzi. Admission control over assured forwarding phbs: a way to provide service accuracy in a diffserv framework. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 4, pages 2561–2565, 2001.

[BBMFP01a] G. Bianchi, N. Blefari-Melazzi, M. Femminella, and F. Pugini. Joint support of qos and mobility in a stateless ip environment. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 6, pages 3454–3458, 2001.

[BBMFP01b] G. Bianchi, N. Blefari-Melazzi, M. Femminella, and F. Pugini. Performance evaluation of a measurement-based algorithm for distributed admission control in a diffserv framework. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 3, pages 1886–1891, 2001.

[BCV01] M. Barry, A. T. Campbell, and A. Veres. Distributed control algorithms for service differentiation in wireless packet networks. In *IEEE INFOCOM*, volume 1, pages 582–590, 2001.

[Bia00] G. Bianchi. Performance analysis of the ieee 802.11 distributed coordination function. In *IEEE JSAC*, volume 18, pages 535–547, March 2000.

[Bla00] D. Black. *Differentiated Services and Tunnels*. IETF RFC 2983, October 2000.

[BMM96] R. Beraldi, S. Marano, and C. Mastroianni. A reversible hierarchical scheme for microcellular systems with overlaying macrocells. In *IEEE INFOCOM'96*, volume 1, pages 51–58, March 1996.

[BZB$^+$97] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. *Resource Reservation Protocol (RSVP)–Version 1 Functional Specification*. IETF RFC 2205, September 1997.

[Cas98] C. Castelluccia. A hierarchical mobility management scheme for ipv6. In *Third Symposium on Computers and Communications (ISCC'98)*, June 1998. Athens, Greece.

[CAV99] G. Chiruvolu, A. Agrawal, and M. Vandenhoute. Mobility and qos support for ipv6-based real-time wireless internet traffic. In *IEEE International Conference on Communications, 1999 (ICC '99)*, volume 1, pages 334–338, 1999.

[CD98] A. Conta and S. Deering. *General Packet Tuneling in IPv6 Specification*. IETF RFC 2473, December 1998.

[CGS$^+$02] A. Campbell, J. Gomez, K. Sanghyo, C-Y Wan, Z.R. Turanyi, and A.G Valko. Comparison of ip micromobility protocols. *IEEE Wireless Communications*, 2002.

[CGW$^+$00] A. Campbell, J. Gomez, C. Wan, S. Kim, Z. Turanyi, and A. Valko. *Cellular IP*. Internet Draft <draft-ietf-mobileip-cellularip-00.txt>, January 2000.

[CH00] Wen-Tsuen Chen and Li-Chi Huang. Rsvp mobility support: a signaling protocol for integrated services internet with mobile hosts. In *IEEE Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, volume 3, pages 1283–1292, 26-30 Mar 2000.

[Cha02] H. Chaskar. *Requirements of a QoS Solution for Mobile IP*. Internet Draft <draft-ietf-mobileip-qos-requirements-03.txt>, July 2002.

[CL95] E. Chlebus and W. Ludwin. Is handoff traffic really poissonian? In *IEEE International Conference on Universal Personal Communications*, pages 348–353, 1995.

[CL00] Shun-Ping Chung and Jin-Chang Lee. Call admission control in cellular multiservice networks using virtual partitioning with priority. In *IEEE International Conference (ICON 2000)*, pages 8–12, 2000.

[CMC$^+$00] J. Chen, A. McAuley, A. Caro, S. Baba, Y. Ohba, and P. Ramanathan. *QoS Architecture Based on Differentiated Services for Next Generation Wireless IP Networks*. Internet Draft <draft-itsumo-wireless-diffserv-00.txt>, July 2000.

[CNP01] M. Carli, A. Neri, and A.R. Picci. Mobile ip and cellular ip integration for inter access networks handoff. In *IEEE International Conference on Communications, 2001 (ICC 2001)*, volume 8, pages 2467–2471, 2001.

[CS00] S. Choi and K. Sohraby. Analysis of a mobile cellular system with hand-off priority and hysteresis control. In *Infocom 2000*, volume 1, pages 217–224, 2000.

[CWKS97] B.P. Crow, I. Widjaja, J.G. Kim, and P. Sakai. Investigation of the ieee 802.11 medium access control (mac) sublayer functions. In *INFOCOM '97*, volume 1, pages 126–133, April 1997.

[CZ02] Yu Cheng and Weihua Zhuang. Diffserv resource allocation for fast handoff in wireless mobile internet. *IEEE Communications Magazine*, 2002.

[DBV$^+$02] R. Droms, J. Bound, B. Volz, T. Lemon, C. Perkins, and M. Carney. *Dynamic Host Configuration Protocol for IPv6 (DHCPv6).* Internet Draft<draft-ietf-dhc-dhcpv6-28.txt>, November 2002.

[DH95] S. Deering and R. Hinden. *Internet Protocol, Version 6 (IPv6) Specification.* IETF RFC 1883, December 1995.

[DHJ$^+$01] S. Deering, B. Haberman, T. Jinmei, E. Nordmark, A. Onoe, and B. Zill. *IPv6 Scoped Address Architecture.* Internet Draft <draft-ietf-ipngwg-scoping-arch-03.txt>, November 2001.

[DKKT00] L. Decreusefond, D. Kofman, H. Korezlioglu, and S. Tohmé. *Eléments de Théorie des Files d'Attente.* Département INFRES, ENST-Paris, 2000.

[DMA00] S. Das, A. Misra, and P. Agrawal. Telemip: Telecommunications-enhanced mobile ip architecture for fast intradomain mobility. *IEEE Personal Communications*, 2000.

[DSAB97] S.K. Das, S.K. Sen, P. Agrawal, and K. Basu. Modelling qos degradation in multimedia wireless networks. In *IEEE International Conference on Personal Wireless Communications*, pages 484 –488, 1997.

[DYP$^+$02] G. Dometty, A. Yegin, C. Perkins, G. Tsirtsis, K. El-Malki, and M. Khalil. *Fast Handovers for Mobile IPv6.* Internet Draft <draft-elmalki-mobileip-fast-handoffs-03.txt>, March 2002.

[EDWS89] S. El-Dolil, W. Wong, and R. Steele. Teletraffic performance of highway microcells with overlay macrocells. In *IEEE JSAC*, volume 7, pages 71–78, 1989.

[EMCH$^+$02] K. El-Malki, P. Calhoun, T. Hiller, J. Kempf, P. McCann, A. Singh, H. Solima, and S. Thalanany. *Low Latency Handoffs in Mobile IPv4.* Internet Draft <draft-ietf-mobileip-lowlatency-handoffs-v4-04.txt>, June 2002.

[EMF99] K. El Malki and N.A. Fikouras. *Fast Handoff Method for Real-Time Traffic over Scalable Mobile IP Networks.* Internet Draft <draft-elmalki-mobileip-fast-handoffs-00.txt>, March 1999.

[EMS00] K. El-Malki and H. Soliman. *Fast Handoffs in Mobile IPv4*. Internet Draft <draft-elmalki-mobileip-fast-handoffs-03.txt>, September 2000.

[ES95] B. Epstein and M. Schwartz. Reservation strategies for multi-media traffic in a wireless environment. In *IEEE 45th Vehicular Technology Conference*, volume 1, pages 165–169, July 1995. Chicago,USA.

[Eur98] European Telecommunications Standards Institute ETSI-Universal Mobile Telecommunications Systems UMTS. *Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS*, April 1998.

[Fit96] P. Fitzpatrick. Performance analysis of a layered wireless network serving different user classes. In *IEEE Vehicular Technology Conference*, pages 431–435, 1996.

[FLW97] P. Fitzpatrick, C.S. Lee, and B. Warfield. Teletraffic performance of mobile radio networks with hierarchical cells and overflow. In *IEEE JSAC*, volume 15, pages 1549–1557, October 1997.

[FRGF96] M. Frullone, G. Riva, P. Grazioso, and G. Falciasecca. Advanced planning criteria for cellular systems. In *IEEE Personal Communications [see also IEEE Wireless Communications]*, volume 3, pages 10–15, Dec 1996.

[GA02] Qiang Gao and A. Acampora. Connection tree based micro-mobility management for ip-centric mobile networks. In *IEEE International Conference on Communications, 2002 (ICC 2002)*, volume 5, pages 3307– 3312, 2002.

[GJP$^+$91] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver Jr., and C.E. Wheatley III. On the capacity of a cellular cdma system. *IEEE Transactions on Vehicular Technology*, 1991.

[GJP02] E. Gustafsson, A. Jonsson, and C. Perkins. *Mobile IPv4 Regional Registration*. Internet Draft <draft-ietf-mobileip-reg-tunnel-07.txt>, October 2002.

[G.P] G.Pujolle. *Tutorial on Resource Allocation in the New Fixed and Mobile Internet Generation*. WATM and Eunice 2001.

[HM93] H. Herzberg and D. McMillan. State-dependent control of call arrivals in layered cellular mobile networks. *Telecommunication Systems*, 1993.

[HM00] H. Haverinen and J.T. Malinen. *Mobile IP Regional Paging (MIRP)*. Internet Draft <draft-haverinen-mobileip-reg-paging-00.txt>, June 2000.

[HR86] D. Hong and S.S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. In *IEEE Transactions On Vehicular Technology*, volume 35, pages 77–92, August 1986.

[IGG93] Chih-Lin I, L. Greenstein, and R. Gitlin. A microcell/macrocell cellular architecture for low- and high-mobility wireless users. In *IEEE JSAC*, volume 11, pages 885–891, 1993.

[ITE] ITEA AMBIENCE. *http://www.extra.research.philips.com/euprojects/ambience/.*

[JF97] B. Jabbari and W.F. Fuhrmann. Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy. In *IEEE JSAC*, volume 15, pages 1539–1548, October 1997.

[JkCC01] Dawi Jeong, Dug kyoo Choi, and Youngjong Cho. The performance analysis of qos provisioning method with buffering and cac in the multimedia wireless internet. In *IEEE 54th Vehicular Technology Conference, 2001 (VTC 2001 Fall)*, volume 2, pages 807–811, 2001.

[JP01] D.B. Jonhson and C. Perkins. *Mobility Support in IPv6.* Internet Draft <draft-ietf-mobileip-ipv6-15.txt>, July 2001.

[KCBN99] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh. Call admission control for adaptive multimedia in wireless/mobile networks. In *ACM workshop on Wireless Mobile Multimedia, WOWMOM'99*, pages 51 –58, 1999.

[KCP00] G. Krishnamurthi, R. Chalmers, and C. Perkins. *Buffer Management for Smooth Handovers in Mobile IPv6.* Internet Draft <draft-krishnamurthi-mobileip-buffer6-00.txt>, July 2000.

[KJ01] J. Kim and A. Jamalipour. Traffic management and qos provisioning in future wireless ip networks. *IEEE Personal Communications*, 2001.

[KK01] Ki-Il Kim and Sang-Ha Kim. Domain based approach for qos provisioning in mobile ip. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 4, pages 2230–2234, 2001.

[KKC00] Kim.S, Kwon.T, and Choi.Y. Call admission control for prioritized adaptive multimedia services in wireless/mobile networks. In *IEEE 51st Vehicular Technology Conference Proceedings*, volume 2, pages 1536–1540, July 2000. Spring Tokyo.

[KKGS99] S. Khurana, A. Kahol, S.K.S. Gupta, and P.K. Srimani. Performance evaluation of distributed co-ordination function for ieee 802.11 wireless lan protocol in presence of mobile and hidden terminals. In *International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 40–47, 1999.

[Kle75] L. Kleinrock. *Queueing Systems, Volume I: Theory.* New York, John Wiley and sons. Wiley-Interscience, 1975.

[Kle76] L. Kleinrock. *Queueing Systems, Volume II: Computer Applications.* New York, John Wiley and sons. Wiley-Interscience, 1976.

[KP00] R. Koodli and C. Perkins. *A Framework for Smooth Handovers with Mobile IPv6.* Internet Draft <draft-koodli-mobileip-smoothv6-00.txt>, July 2000.

[KSK01] Kanghee. Kim, Seokjoo Shin, and Kiseon Kim. A novel mac scheme for prioritized services in ieee 802.11a wireless lan. In *IEEE International Symposium on ATM (ICATM 2001) and High Speed Intelligent Internet Symposium*, pages 196–199, 2001.

[KT95] L. Kleinrock and F.A. Tobagi. Packet switching in radio channels: Part 2: the hidden terminal problem in carrier sense multiple-access and the busy-tone solution. In *IEEE Transactions on Communications COM-23*, 1995.

[KW01] A. Kospel and A. Wolisz. Voice transmission in an ieee 802.11 wlan based access network. In *WoWMoM*, pages 24–33, July 2001. Rome, Italy.

[KZZM01] Zhigang Kan, Dongmei Zhang, Runtong Zhang, and Jian Ma. Qos in mobile ipv6. In *International Conferences on Info-tech and Info-net, 2001 (ICII 2001)*, volume 2, pages 492–497, 2001.

[LAN97] LAN MAN Standards of the IEEE Computer Society.IEEE Standard 802.11. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, 1997. http://www.ieee802.org/11/.

[LDK91] M. Law, Averill and W. David Kelton. *Simulation Modeling and Analysis*. McGRAW HILL International Editions, 1991.

[LG95] X. Lagrange and P. Godlewski. Teletraffic analysis of a hierarchical cellular network. In *IEEE Vehicular Technology Conference*, volume 2, pages 882–886, July 1995.

[LMN94a] Y. Lin, S. Mohan, and A. Noerpel. Analyzing the trade off between implementation costs and performance: Pcs channel assignment strategies for hand-off and initial access. *IEEE Personal Communications*, 1994.

[LMN94b] Y. Lin, S. Mohan, and A. Noerpel. Queueing priority channel assignment strategies for pcs hand-off and initial access. In *IEEE Transactions On Vehicular Technology*, volume 43, pages 704–712, August 1994.

[LNH94] Y. Lin, A. Noerpel, and D. Harasty. A non-blocking channel assignment strategy for hand-offs. In *IEEE Third Annual Conference on Universal Personal Communications ICUPC*, pages 558–562, 1994.

[LNH96] Y. Lin, A. Noerpel, and D. Harasty. The sub-rating channel assignment strategy for pcs hand-offs. In *IEEE Transactions on Vehicular Technology*, volume 45, pages 122–129, February 1996.

[LZMF98] B. Li, Q. Zeng, K. Mukumoto, and A. Fukuda. A preemptive priority handover scheme in integrated voice and data cellular mobile systems. In *International Conference on Communication Technology (ICCT 1998)*, volume 1, pages 67–71, 1998.

[MAB01] J. McNair, I.F. Akyildiz, and M.D. Bender. Handoffs for real-time traffic in mobile ip version 6 networks. In *IEEE GLOBECOM '01*, volume 6, pages 3463–3467, 2001.

[MC01] A. Mohammad and A. Chen. Seamless mobility requirements and mobility architectures. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 3, pages 1950–1956, 2001.

[McM95] D. McMillan. Delay analysis of a cellular mobile priority queuing system. *IEEE/ACM Transactions on Networking*, 1995.

[MDM+00] A. Misra, S. Das, A. Mcauley, A. Dutta, and S.K. Das. Integrating qos support in telemip's mobility architecture. In *IEEE International Conference on Personal Wireless Communications, 2000*, pages 57–64, 2000.

[MH99] A. Mahmoodian and G. Haring. A resource allocation mechanism to provide guaranteed service to mobile multimedia applications. In *First IEEE/Popov Workshop on Internet Technologies and Services 1999*, pages 9–17, 1999.

[MLP01] J.T. Malinen, F. Le, and C. Perkins. *Mobile IPv6 Regional Registration.* Internet Draft <draft-malinen-mobileip-regreg6-01.txt>, March 2001.

[MWM+97] M. Madfors, K. Wallstedt, S. Magnusson, H. Olofsson, P.-O. Backman, and S. Engstrom. High capacity with limited spectrum in cellular systems. *IEEE Communications Magazine*, 1997.

[NA95] M. Naghshineh and A.S. Acampora. Qos provisioning in micro-cellular networks supporting multimedia traffic. In *INFOCOM '95*, volume 3, pages 1075–1084, 1995.

[NBBB98] K. Nichols, S. Blake, F. Baker, and D. Blake. *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers.* IETF RFC 2474, December 1998.

[Net] Network Simulator, ns-2. *http://www.isi.edu/nsnam/ns/.*

[NNS98] T. Narten, E. Nordmark, and W. Simpson. *Neighbor Discovery for IP Version 6 (IPv6).* IETF RFC 1883, December 1998.

[NS96] M. Naghshineh and M. Schwartz. Distributed call admission control in mobile/wireless networks. *IEEE Journal on Selected Areas in Communications*, 1996.

[OGA99] L. Ortigoza-Guerrero and A.Hamid Aghvami. A prioritized handover dynamic channel allocation strategy for pcs. In *IEEE Transactions On Vehicular Technology*, volume 48, pages 1203–1215, July 1999.

[OGA00] L. Ortigoza-Guerrero and A.H. Aghvami. *Resource Allocation in Hierarchical Cellular Systems.* Artech House Publishers, 2000.

[OKS98] C. Oliveira, J.B. Kim, and T. Suda. An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks. *IEEE Journal on Selected Areas in Communications*, 1998.

[OMN] OMNeT++. *http://www.hit.bme.hu/phd/vargaa/omnetpp.htm.*

[Per96a] C. Perkins. *IP Encapsulation within IP.* IETF RFC 2003, May 1996.

[Per96b] C. Perkins. *IP Mobility Support.* IETF RFC 2002, October 1996.

[PKZ01] S. Paskalis, A. Kaloxylos, and E.E. Zervas. An efficient qos scheme for mobile hosts. In *26th Annual IEEE Conference on Local Computer Networks, 2001 (LCN 2001)*, pages 630–637, 2001.

[Pro99] Projet MINICELL. *Livrable 2: Description of UMTS Environment*, October 1999.

[Pro00] Projet MINICELL. *Livrable 6: Specification of the UMTS Environment Simulator*, Mars 2000.

[QC01] Daji Qiao and Sunghyun Choi. Goodput enhancement of ieee 802.11a wireless lan via link adaptation. In *EEE International Conference on Communications ICC*, volume 7, pages 1995–2000, 2001.

[RB01] P. Reinbold and O. Bonaventure. A comparison of ip mobility protocols. Technical report, University of Namur, 2001.

[Rey00] P. Reynolds. Mobility management for the support of handover within a heterogeneous mobile environment. In *IEE International Conference on 3G Mobile Communication Technologies*, pages 341–346, 2000.

[RH94] S.S. Rappaport and L. Hu. Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis. In *Proceedings of the IEEE*, volume 82, pages 1383–1397, 1994.

[RH95] S.S. Rappaport and L. Hu. Personal communication systems using multiple hierarchical cellular overlays. In *IEEE JSAC*, volume 13, pages 406–414, 1995.

[RLPST$^+$99] R. Ramjee, T. La Porta, S. S. Thuel, K. Vardhan, and S.Y. Wang. Hawaii: A domain-based approach for supporting mobility in wide-area wireless networks. In *Seventh International Conference on Network Protocols (ICNP '99)*, pages 283–292, November 1999. Toronto, Canada.

[RLPTV99] R. Ramjee, T. La Porta, S. Thuel, and K. Vardhan. *IP Micro-Mobility Support Through HAWAII*. Internet Draft <draft-ramjee-micro-mobility-hawaii-00.txt>, September 1999.

[SCEMB01] H. Soliman, C. Castellucia, K. El-Malki, and L. Bellier. *Hierarchical MIPv6 Mobility Management (HMIPv6)*. Internet Draft <draft-ietf-mobileip-hmipv6-05.txt>, July 2001.

[SGCW00] Z.D. Shelby, D. Gatzounas, A. Campbell, and C. Wan. *Cellular IPv6*. Internet Draft <draft-shelby-seamoby-cellularipv6-00.txt>, November 2000.

[SMG$^+$01] Z.D. Shelby, P. Mahonen, D. Gatzounas, A. Inzerilli, and V. Typpo. *Cellular IP Route Optimization*. Internet Draft <draft-shelby-cip-routeoptimization-00.txt>, June 2001.

[SN92] R. Steele and M. Nofal. Teletraffic performance of microcellular personal communication networks. In *Communications, Speech and Vision, IEE Proceedings I*, volume 139, pages 448–461, 1992.

[SPG97] S. Shenker, C. Patridge, and R. Guerin. *Specification of Guaranteed Quality of Service.* RFC 2212, September 1997.

[SSL01] Q. Shen, Q. Seah, and Lo.A. *Flow Transparent Mobility and QoS Qupport for IPv6-based Wireless Real-time Services.* Internet Draft <draft-shen-ipv6-flow-trans-00.txt>, Feb. 2001.

[ST99] T. Suzuki and S. Tasaka. Performance evaluation of integrated video and data transmission with the ieee 802.11 standard mac protocol. In *GLOBECOM '99*, volume 1B, pages 580 –586, 1999.

[ST01] T. Suzuki and S. Tasaka. Performance evaluation of priority-based multimedia transmission with the pcf in an ieee 802.11 standard wireless lan. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, volume 2, pages G–70–G–77, 2001.

[TBA01] A. Talukdar, B.R. Badrinath, and A. Acharya. Mrsvp: A resource reservation protocol for an integrated services network with mobile hosts. *Wireless Networks*, 2001.

[TBC99] C.H. Tse, B. Bensaou, and K.C. Chua. Efficient distributed scheduling architecture for wireless atm networks. In *ATM Workshop, 1999. IEEE Proceedings*, pages 415 –425, 1999.

[TJ92] S. Tekinay and B. Jabbari. A measurement-based prioritization scheme for handovers in mobile cellular networks. *IEEE Journal on Selected Areas in Communications*, 1992.

[TLL01] Chien-Chao Tseng, Gwo-Chuan Lee, and Ren-Shiou Liu. Hmrsvp: a hierarchical mobile rsvp protocol. In *International Conference on Distributed Computing Systems Workshop, 2001*, pages 467–472, Apr 2001.

[TN98] S. Thomson and T. Narten. *IPv6 Stateless Address Autoconfiguration.* IETF RFC 2462, December 1998.

[TSZ99] A. Terzis, M. Srivastava, and Lixia Zhang. A simple qos signaling protocol for mobile hosts in the integrated services internet. In *Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, volume 3, pages 1011–1018, 21-25 Mar 1999.

[TWH99] Yueh-Eo Tseng, Hsiao-Kuang Wu, and Ming-I Hsieh. Connection admission control for qos guarantees in mobile networks. In *Eight International Conference on Computer Communications and Networks, 1999*, pages 542–547, 1999.

[VCBS01] A. Veres, A. T. Campbell, M. Barry, and Li-Hsiang Sun. Supporting service differentiation in wireless packet networks using distributed control. In *IEEE JSAC*, volume 19, pages 2081 –2093, Oct 2001.

[Wro97a] J. Wroclawski. *Specification of the Controlled-Load Network Element Service.* RFC 2211, September 1997.

[Wro97b] J. Wroclawski. *The Use of RSVP with IETF Integrated Services.* IETF RFC 2210, September 1997.

[WWL98] S. Wu, K.Y.M. Wong, and B. Li. A new distributed and dynamic call admission policy for mobile wireless networks with qos guarantee. In *IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, volume 1, pages 260–264, 1998.

[XPC01] Y. Xiao and C.L. Philip Chen. Qos for adaptive multimedia in wireless/mobile networks. In *Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 81–88, 2001.

[XPCW00] Y. Xiao, C.L. Philip Chen, and Y. Wang. Quality of service and call admission control for adaptive multimedia services in wireless/mobile networks. In *IEEE 2000 National Aerospace and Electronics Conference NAECON*, pages 214 –220, 2000.

[XPCW01] Y. Xiao, C.L. Philip Chen, and Y. Wang. Fair bandwidth allocation for multi-class of adaptive multimedia services in wireless/mobile networks. In *IEEE 53rd Vehicular Technology Conference VTC*, volume 3, pages 2081 –2085, 2001.

[YLLK00] Suk-Un Yoon, Ji-Hoon Lee, Ki-Sun Lee, and Chul-Hee Kang. Qos support in mobile/wireless ip networks using differentiated services and fast handoff method. In *IEEE Wireless Communications and Networking Conference, 2000 (WCNC 2000)*, volume 1, pages 266–270, 2000.

[YNH01] S. Yasukawa, J. Nishikido, and K. Hisashi. Scalable mobility and qos support mechanism for ipv6-based real-time wireless internet traffic. In *IEEE Global Telecommunications Conference, 2001 (GLOBECOM '01)*, volume 6, pages 3459–3462, 2001.

[ZA00] Q. Zeng and D.P Agrawal. Performance analysis of a handover scheme in integrated voice/data wireless networks. In *IEEE Vehicular Technology Conference, Fall VTC*, volume 4, pages 1986–1992, 2000.

[ZDE$^+$02] Lixia Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. Rsvp: a new resource reservation protocol. *IEEE Communications Magazine*, 2002.

[ZK01] J. Zander and S. Kim. *Radio Resource Management For Wireless Networks.* Artech House Publishers, 2001. Software authors: M.Almgren and O.Queseth.

[ZKC96] D. Zeghlache, F. Khan, and D. Calin. On the performance of high to low level hand-off in a hierarchical personal communication system. In *ACTS Mobile Telecommunication Summit*, pages 218–223, 1996.

List of Publications

- R. Naja and S. Tohmé. *Centralized And Distributed Adaptive Call Admission Control in Cellular IP Networks.* In the 14th IEEE International Symposium On Personal, Indoor And Mobile Radio Communications PIMRC 2003, September 2003, Beijing, China.
- Y.M. Ghamri-Doudane, R. Naja, G. Pujolle and S. Tohmé. *P3-DCF: Service Differentiation in IEEE 802.11 WLANs using Per-Packet Priorities.* In the IEEE Vehicular Technology Conference VTC Fall 2003, October 2003, Orlando, US.
- R. Naja and S. Tohmé. *Smooth Handover and Optimized Uplink Routing in Cellular IPv6 Networks.* In Personal Wireless Communications Conference PWC 2003, September 2003, Venise, Italy.
- R. Naja and S. Tohmé. *QoS Provisioning in Mobile Wireless Multimedia Networks Using a Dynamic Adaptive Architecture : DYNAA.* In Personal Wireless Communications Conference PWC 2002, October 2002, Singapore.
- R. Naja and S. Tohmé. *DYNAA: Dynamic Adaptive Architecture for Multimedia Services Wireless Mobile Multimedia Networks.* In IFIP WG6.7 Workshop and Eunice Summer School on Adaptable Networks and Teleservices, September 2002, Trondheim, Norway.
- R. Naja and S. Tohmé. *QoS Provisioning and Handover Issues in Mobile Wireless Multimedia Networks.* In IEEE Workshop on Applications and Services in Wireless Networks ASWN (ASWN 2002), July 2002, Paris.
- R. Naja and S. Tohmé. *Multi-Service Call Admission Control and Handover Issues in Wireless Multimedia Networks.* Mobiles-services et réseaux mobiles de 3ème Génération MS3G 2001, December 2001, Lyon.

Technical Reports

- ALL AMBIENCE members. *Ambience WP1 System Architecture*, ITEA Ambience Project, Delivrable D1.2, July 2002.
- R. Naja et al. *QoS et Performances dans l'UTRAN.* Minicel Project, Delivrable 8, November 2000.

Appendix A

Simulation Environment and Tools

The simulation approach is currently the most reasonable refuge for researchers to evaluate the performance of real complex systems. Analytical solutions to mathematical models are not always available to investigate the behavior of some systems. Moreover, assumptions that may be considered to simplify the mathematical model can sometimes provide misleading results as a price for reduced model complexity.

In contrast to analytical evaluation, which can be acceptably valid for simple procedures, simulation permits a relatively faithful representation of the behavior of larger and more complex systems. In addition, simulation analysis provides more flexibility in testing the system under study for varying parameters and environments. An optimization of the system performance can then be realized.

The simulation approach is the main technique used to study and analyze the behavior of the multiple mechanisms proposed in this thesis. Simulation was also used in order to validate the mathematical model elaborated for the multi-service CAC proposed in chapter 4. Simulation models were built for each studied scheme, taking into consideration the smallest details of its operation. The proposed mechanisms are developed using the *Network Simulator ns* and the *OMNeT++*, whose features are briefly described in the next sections.

A.1 Network Simulator *ns*

The Network Simulator *ns* is an event-driven simulator. It is developed in the context of the VINT project as a collaboration between researchers at UC Berkeley, LBL, USC/ISI and Xerox PARC.

ns is an object oriented simulator, written in C++, with an OTcl interpreter as a front-end. The simulator supports the class hierarchy in C++, called the compiled class and a similar class hierarchy within the OTcl, called the interpreted class. The two hierarchies are closely related to each other; from the user's perspective, there is a one-to-one correspondence between a class in the interpreted hierarchy and its counterpart in the compiled hierarchy.

C++ is fast to run but slower to change, making it suitable for detailed protocol implementation. In contrast, OTcl runs much slower but can be changed very quickly and interactively, making it more suitable for simulation configuration. One of the major difficulties of *ns* is to decide which part of the program should be implemented in which language.

Finally *ns* is an open source software. For more information about the simulator, the interested reader might want to have look at [Net].

A.2 *OMNeT++*

OMNeT++ stands for Objective Modular Network Testbed in C++ [OMN]. *OMNeT++* is an object-oriented modular discrete event simulator. It can be used for modeling communication protocols, computer networks and traffic modeling, multi-processor and distributed systems, administrative systems, ... and any other system where the discrete event approach is suitable.

An *OMNeT++* model consists of hierarchically nested modules. Modules communicate with message passing. Messages can contain arbitrarily complex data structures. Modules can send messages either directly to their destination or along a predefined path, through gates and connections. Modules at the lowest level of the module hierarchy are to be provided by the user, and they contain the algorithms in the model.

A.3 Output Analysis

Several performance measures were collected from the simulated models, by means of different types of probes. These probes extract the necessary information from the field values marked in the data structures, while flowing through the system. Statistical significance of the collected measures can be estimated using the confidence interval technique.

A.3.1 Confidence interval calculation

Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed measured values, with finite mean μ and finite variance σ^2, then the sample mean is:

$$\overline{X}(n) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{A.1}$$

and the sample variance is:

$$S^2(n) = \frac{\sum_{i=1}^{n} [X_i - \overline{X}(n)]^2}{n-1} \tag{A.2}$$

If n is sufficiently large, then using the central limit theorem, we can consider that the sample mean $\overline{X}(n)$ is approximately distributed as a normal random variable with mean μ and variance σ^2/n. The $100(1-\alpha)$ percent confidence interval for μ can be calculated from the formula:

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}} \tag{A.3}$$

where

$$t_n = \frac{\overline{X}(n) - \mu}{\sqrt{S^2(n)/n}} \tag{A.4}$$

has a t distribution with $n-1$ degrees of freedom. In the expression A.3, $t_{n-1,1-\alpha/2}$ is the upper $1-\alpha/2$ critical point for the t distribution [LDK91].

A.3.2 Confidence intervals technique

There are three commonly applied techniques for obtaining confidence intervals of simulation results.

- Independent replication of the simulation: where independent X_is are obtained by iterating the same simulation over the seed parameter of the random number generators, to provide several independent simulation runs.
- Batch means: in which a single long simulation run is broken into n batches. The duration of each batch should be long enough, such that the means from successive batches can be assumed independent.
- Regenerative sampling: which is mainly based on the definition of regenerative states in a queuing model. Any stable and non-degenerate system will visit a given state repeatedly, where this state is defined as regeneration point. The time durations between visits are called regeneration cycles. Measures are then collected for regenerative cycles which are generally statistically independent. The regenerative sampling method is practically more difficult to apply than the first two methods, especially in complex systems.

In our work, the independent replication method was used. Thus, several runs have been performed for each simulated model, with different values of random number generator seeds, to find a 95% confidence interval in the collected system statistics for some simulations.

Simulations were also repeated for variable time durations, until a stability is attained, in the evaluated results and in the simulation sequence of events. The simulation time was determined such that a certain number of events are generated within the simulation run to obtain a given order of magnitude for the results.