ΚΑΤΑΝΟΜΗ ΡΟΩΝ ΚΙΝΗΣΗΣ ΣΕ ΑΣΥΡΜΑΤΟΥΣ ΣΤΑΘΜΟΥΣ ΜΕ ΠΟΛΛΑΠΛΕΣ ΕΝΕΡΓΕΣ ΔΙΚΤΥΑΚΕΣ ΔΙΕΠΑΦΕΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Βασίλειος Η. Ζαφείρης
Φεβρουάριος 2011

TRAFFIC FLOW ASSIGNMENT FOR MULTI-HOMED WIRELESS HOSTS
SYSTEM ARCHITECTURE AND ALGORITHMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF INFORMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Vassilis E. Zafeiris
February 2011

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΚΑΤΑΝΟΜΗ ΡΟΩΝ ΚΙΝΗΣΗΣ ΣΕ ΑΣΥΡΜΑΤΟΥΣ ΣΤΑΘΜΟΥΣ ΜΕ ΠΟΛΛΑΠΛΕΣ ΕΝΕΡΓΕΣ ΔΙΚΤΥΑΚΕΣ ΔΙΕΠΑΦΕΣ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: ΒΑΣΙΛΕΙΟΣ Η. ΖΑΦΕΙΡΗΣ
ΕΠΙΒΛΕΠΩΝ: ΕΜΜΑΝΟΥΗΛ ΓΙΑΚΟΥΜΑΚΗΣ

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ
ΕΜΜΑΝΟΥΗΛ ΓΙΑΚΟΥΜΑΚΗΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.
ΝΙΚΟΛΑΟΣ ΜΑΛΕΥΡΗΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.
ΓΕΩΡΓΙΟΣ ΞΥΛΩΜΕΝΟΣ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

ΜΙΛΤΙΑΔΗΣ ΑΝΑΓΝΩΣΤΟΥ, ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

ΕΜΜΑΝΟΥΗΛ ΓΙΑΚΟΥΜΑΚΗΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΕΥΑΓΓΕΛΟΣ ΜΑΓΕΙΡΟΥ, ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΝΙΚΟΛΑΟΣ ΜΑΛΕΥΡΗΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΛΑΖΑΡΟΣ ΜΕΡΑΚΟΣ, ΚΑΘΗΓΗΤΗΣ Ε.Κ.Π.Α.

ΓΕΩΡΓΙΟΣ ΞΥΛΩΜΕΝΟΣ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΓΕΩΡΓΙΟΣ ΠΟΛΥΖΟΣ, ΚΑΘΗΓΗΤΗΣ Ο.Π.Α.

ΗΜΕΡΟΜΗΝΙΑ ΕΞΕΤΑΣΗΣ
7 ΦΕΒΡΟΥΑΡΙΟΥ 2011

Αφιέρωση

Αφιερώνεται στον πατέρα μου, Ηλία.

Περίληψη

Η διατριβή στοχεύει στη βελτιστοποίηση της συνδεσιμότητας ενός κινητού τερματικού σε ένα ετερογενές περιβάλλον ασύρματης δικτυακής πρόσβασης. Το δικτυακό αυτό περιβάλλον, που συνήθως αναφέρεται ως 4G, χαρακτηρίζεται από πολλαπλά δίκτυα ασύρματης πρόσβασης, πιθανό διαφορετικής τεχνολογίας, που διασυνδέονται και διαλειτουργούν μέσω μιας IP δικτυακής υποδομής. Στα πλαίσια της διατριβής γίνεται μελέτη του προβλήματος της κατανομής των ροών κίνησης ενός ασύρματου σταθμού που διαθέτει δυνατότητα ταυτόχρονης χρήσης πολλαπλών δικτυακών διεπαφών (multi-homed). Επιπλέον, γίνεται προδιαγραφή της αρχιτεκτονικής και μελέτη της επίδοσης ενός συστήματος για την υποστήριξη της εκτέλεσης αλγορίθμων κατανομής ροών κίνησης ή συναφών μηχανισμών απόφασης, όπως η επιλογή δικτύου πρόσβασης.

Το πρόβλημα της κατανομής ροών κίνησης (traffic flow assignment problem- TFAP) αποτελεί προσαρμογή για multi-homed σταθμούς του προβλήματος επιλογής δικτύου πρόσβασης. Η επιλογή δικτύου πρόσβασης αποτελεί τμήμα του μηχανισμού διαχείρισης διαπομπών ενός κινητού τερματικού και στοχεύει στην επιλογή του καλύτερου σημείου και υπηρεσίας ασύρματης πρόσβασης για την διατήρηση της συνδεσιμότητάς του. Σε ένα ετερογενές περιβάλλον ασύρματης πρόσβασης, το εύρος των εναλλακτικών επιλογών επιτρέπει στο μηχανισμό επιλογής δικτύου την εξυπηρέτηση επιπρόσθετων στόχων του χρήστη πέραν της διατήρησης συνδεσιμότητας, π.χ. εξοικονόμηση χρημάτων και ενέργειας. Επιπλέον, η επιλογή δικτύου πρόσβασης σε ένα multi-homed σταθμό (τερματικό ή κινητό δρομολογητή) προϋποθέτει αποφάσεις σχετικά με την επιλογή των δικτυακών διεπαφών που θα πρέπει να ενεργοποιηθούν καθώς και την κατανομή των ροών κίνησης σε αυτές. Συνεπώς, το πρόβλημα επιλογής δικτύου στο πλαίσιο αυτό συνίσταται από τρία υπο-προβλήματα: (α) επιλογή δικτυακών διεπαφών προς ενεργοποίηση, (β) επιλογή υπηρεσιών πρόσβασης για τις ενεργές δικτυακές διεπαφές, (γ) κατανομή των ροών κίνησης στις ενεργές δικτυακές διεπαφές.

Η διατριβή χειρίζεται, στο Κεφάλαιο 3, τα τρία υπό-προβλήματα με ενιαίο τρόπο μέσω

της διατύπωσης του προβλήματος κατανομής ροών κίνησης (traffic flow assignment problem-TFAP). Το πρόβλημα στοχεύει στην κατανομή των ροών κίνησης (εισερχομένων ή εξερχομένων) των εφαρμογών σε κατάλληλες δικτυακές διεπαφές και υπηρεσίες μεταφοράς δεδομένων με τρόπο που να εξασφαλίζει τον καλύτερο συνδυασμό οικονομικού κόστους και κατανάλωσης ενέργειας. Το οικονομικό κόστος αναφέρεται στο κόστος χρήσης των δικτύων, ενώ η κατανάλωση ενέργειας οφείλεται στη λειτουργία των ενεργών δικτυακών διεπαφών. Το αντιστάθμισμα που υπάρχει μεταξύ των δυο παραγόντων κόστους βασίζεται στις εξής παραδοχές: (α) διαθεσιμότητα υπηρεσιών μεταφοράς δεδομένων συγκρίσιμου κόστους στις διάφορες δικτυακές διεπαφές και (β) ύπαρξη συσχέτισης μεταξύ κόστους χρήσης και προσφερόμενης χωρητικότητας από τις διάφορες υπηρεσίες. Βάσει των παραπάνω παραδοχών και δεδομένου ότι ο φόρτος κίνησης του τερματικού δεν μπορεί να εξυπηρετηθεί αποκλειστικά από μια δικτυακή διεπαφή (λόγω περιορισμών χωρητικότητας ή ποιότητας υπηρεσίας), η ελαχιστοποίηση του οικονομικού κόστους περιλαμβάνει την κατανομή της κίνησης σε δικτυακές διεπαφές με πρόσβαση στις φθηνότερες υπηρεσίες μεταφοράς δεδομένων. Αντίστοιχα, η ελαχιστοποίηση της καταναλισκόμενης ισχύος απαιτεί την ενεργοποίηση του μικρότερου δυνατού αριθμού δικτυακών διεπαφών, μέσω της χρήσης υπηρεσιών υψηλής χωρητικότητας και συνήθως υψηλότερου κόστους.

Η προτεινόμενη στη διατριβή μαθηματική διατύπωση του TFAP καταλήγει σε ένα πρόβλημα συνδυαστικής βελτιστοποίησης δύο στόχων. Για την επίλυση του προβλήματος γίνεται αρχικά μετατροπή του σε πρόβλημα βελτιστοποίησης ενός στόχου. Συγκεκριμένα, το οικονομικό κόστος επιλέγεται ως βασικός στόχος προς βελτιστοποίηση, ενώ η κατανάλωση ενέργειας προστίθεται σαν επιπλέον περιορισμός του προβλήματος μέσω του ορισμού ενός άνω ορίου για τις επιτρεπόμενες τιμές του. Το συγκεκριμένο άνω όριο δεν είναι σταθερό για όλα τα στιγμιότυπα του TFAP, αλλά εξαρτάται από την κατάσταση της κινητής συσκευής (πχ., επίπεδο φόρτισης της μπαταρίας), το περιβάλλον λειτουργίας της και υπολογίζεται από το υποσύστημα διαχείρισης ενέργειας της συσκευής. Στα πλαίσια της διατριβής γίνεται μελέτη της πολυπλοκότητας επίλυσης του TFAP μέσω αναγωγής από το πρόβλημα Πολλαπλών Σακιδίων με Περιορισμούς Ανάθεσης (Multiple Knapsack Problem with Assignment Restrictions- MKAR). Καθώς το MKAR είναι NP-Hard, το TFAP έχει αντίστοιχη πολυπλοκότητα και, συνεπώς, απαιτούνται προσεγγιστικοί αλγόριθμοι για την γρήγορη παραγωγή λύσεών του.

Για το σκοπό αυτό, γίνεται σχεδιασμός ενός ευρετικού αλγορίθμου βασισμένου σε τοπική αναζήτηση. Ο αλγόριθμος χαρακτηρίζεται από αποδοτικούς χρόνους εκτέλεσης για ένα ευρύ φάσμα προβλημάτων ρεαλιστικού μεγέθους. Η ποιότητα της προσέγγισης των πραγματικών λύσεων κρίνεται ικανοποιητική μέσω αξιολόγησης που βασίστηκε στη σύγκριση ευρετικών και πραγματικών λύσεων για ένα μεγάλο σύνολο τυχαία παραγομένων προβλημάτων. Συγκεκριμένα, το σφάλμα προσέγγισης του αλγορίθμου όσον αφορά το οικονομικό κόστος δεν ξεπερνά κατά μέσο όρο το 8,1% για προβλήματα μικρού και μεσαίου μεγέθους. Όσον αφορά την κατανομή του σφάλματος προσέγγισης, το 80% των παραπάνω προβλημάτων επιλύθηκαν με σφάλμα μικρότερο του 15% ενώ το 95% αυτών με σφάλμα μικρότερο του 35%. Καθώς η παραγωγή πραγματικών λύσεων για μεγάλα προβλήματα αποτελεί χρονοβόρα διαδικασία, η αξιολόγηση της ποιότητας λύσεων του αλγορίθμου για τέτοιου είδους προβλήματα βασίστηκε στις λύσεις μιας χαλαρωμένης εκδοχής του TFAP. Το χαλαρωμένο πρόβλημα προκύπτει από το TFAP μέσω της αφαίρεσης του περιορισμού ακεραιότητας ως προ την κατανομή των ροών, δηλαδή επιτρέποντας την διάσπαση της κίνησης μεμονωμένων ροών σε δύο ή περισσότερες δικτυακές διεπαφές. Οι λύσεις του χαλαρωμένου προβλήματος είναι όμοιες ή καλύτερες των αντίστοιχων του αρχικού προβλήματος. Συνεπώς, το σφάλμα προσέγγισης είναι υπερεκτιμημένο. Παρά το γεγονός αυτό, το μέσο σφάλμα προσέγγισης κατά την επίλυση προβλημάτων μεγάλου μεγέθους δεν ξεπερνά το 13,1%. Όσον αφορά την κατανομή του σφάλματος, το 80% των μεγάλων προβλημάτων επιλύονται με ακρίβεια μεγαλύτερη του 20%, ενώ το 90% αυτών με ακρίβεια μεγαλύτερη του 29%.

Η διαπίστωση των ωφελειών, στο πεδίο του χρόνου, της βελτιστοποιημένης κατανομής κίνησης, καθώς και η εκτίμηση της επιβάρυνσης που προκαλείται όσον αφορά τη διαχείριση κινητικότητας, αξιολογήθηκαν στην παρούσα διατριβή μέσω προσομοίωσης. Το σύστημα προσομοίωσης, που υλοποιήθηκε για το σκοπό αυτό σε Java, είναι βασισμένο στην παραγωγή και επεξεργασία τυχαίων συμβάντων. Το σύστημα προσομοιώνει την λειτουργία ενός κινητού δρομολογητή, για χρονική διάρκεια τριών ωρών, καθώς εξυπηρετεί 5 χρήστες σε ένα δικτυακό περιβάλλον που αποτελείται από 4 UMTS και 10 WLAN δίκτυα. Ο δρομολογητής διαθέτει πρόσβαση στις υπηρεσίες των παραπάνω δικτύων μέσω 2 UMTS και 2 WLAN δικτυακών διεπαφών. Κάθε χρήστης παράγει, στη διάρκεια της προσομοίωσης, συνεδρίες τηλεδιάσκεψης και μεταφοράς αρχείων μέσω HTTP/FTP σύμφωνα με κοινώς αποδεκτά μοντέλα κίνησης. Το σύστημα προσομοιώνει, επίσης, την προσαρμογή του κινητού δρομολογητή στις διαρκώς μεταβαλλόμενες συνθήκες δικτύου και απαιτήσεων μεταφοράς δεδομένων, μέσω της επαναληπτικής επίλυσης TFAP στιγμιοτύπων. Τα αποτελέσματα από την εκτέλεση 100 προσομοιώσεων δείχνουν μια αύξηση στο οικονομικό κόστος ίση με 7,5%, κατά μέσο όρο, σε σχέση με το βέλτιστο συνολικό κόστος, όταν χρησιμοποιείται ο προτεινόμενος ευρετικός αλγόριθμος. Από την άλλη πλευρά, η χρήση ενός εναλλακτικού αλγορίθμου, που έχει προταθεί στη βιβλιογραφία, έχει ως αποτέλεσμα αύξηση 17% στο μέσο οικονομικό κόστος σε σχέση με το βέλτιστο. Η επιβάρυνση που προκύπτει ως προς τη διαχείριση κινητικότητας, λόγω της προσαρμογής της κατάστασης του δρομολογητή βάσει των λύσεων των TFAP προβλημάτων, ισούται με 2,2 μετακινήσεις ροών και 1,5 οριζόντιες διαπομπές ανά λεπτό κατά μέσο όρο. Η επιβάρυνση αυτή είναι ανεκτή, δεδομένου του αριθμού χρηστών που εξυπηρετούνται και της ωφέλειας από οικονομικής άποψης. Τα αποτελέσματα αυτά συντείνουν στην πρακτικότητα εφαρμογής της προσέγγισης.

Η αξιοποίηση αλγορίθμων έξυπνης επιλογής δικτύου ή κατανομής ροών κίνησης δεν μπορεί να βασιστεί αποκλειστικά σε υποδομή εγκατεστημένη στα κινητά τερματικά. Οι παραπάνω μηχανισμοί απόφασης απαιτούν τη χρήση πληροφορίας η οποία είναι διαθέσιμη, τόσο τοπικά στα τερματικά (πχ., απαιτήσεις μεταφοράς δεδομένων, προτιμήσεις χρηστών), όσο και γεωγραφικά εξαρτώμενης πληροφορίας που αφορά τη διαθεσιμότητα δικτύων και υπηρεσιών. Η αναζήτηση της τελευταίας αποκλειστικά μέσω ανίχνευσης από τις δικτυακές διεπαφές του τερματικού αποτελεί χρονοβόρα και ενεργοβόρα διαδικασία. Επιπλέον, η έγκαιρη και αξιόπιστη λήψη πληροφορίας, σχετικά με τη διαθεσιμότητα δικτύων πρόσβασης, αποκλειστικά από ένα δικτυακό πάροχο για κάθε χρήστη (πχ., από τον οικείο πάροχό του), πολλές φορές δεν είναι εφικτή. Ο κύριο λόγος είναι η έλλειψη κινήτρων του παρόχου για προώθηση των υπηρεσιών των ανταγωνιστών του.

Στο Κεφάλαιο 4 της διατριβής προτείνεται η αρχιτεκτονική ενός συστήματος για την υποστήριξη της εκτέλεσης αλγορίθμων επιλογής δικτύου (ή διαφορετικά απόφασης διαπομπής) και κατανομής ροών κίνησης. Το σύστημα είναι κατανεμημένο σε πολλαπλές διαχειριστικές περιοχές και είναι βασισμένο σε πράκτορες λογισμικού. Οι πράκτορες φιλοξενούνται σε πλατφόρμες εκτέλεσης πρακτόρων οι οποίες βρίσκονται εγκατεστημένες στις διάφορες διαχειριστικές περιοχές. Οι πράκτορες λογισμικού εκπροσωπούν τους χρήστες, τους δικτυακούς παρόχους, ένα Πάροχο Πολλαπλής Πρόσβασης (Multi-Access Provider-MAP) και τη ρυθμιστική αρχή. Ο MAP αποτελεί μια επιχειρηματική οντότητα που διατηρεί συμβόλαια περιαγωγής με τους δικτυακούς παρόχους και επιτρέπει την πρόσβαση στις υπηρεσίες τους μέσω μιας ενιαίας συνδρομής από το χρήστη. Επιπλέον, ο MAP προσφέρει, στα πλαίσια του συστήματος, υπηρεσίες διαπίστευσης, εξουσιοδότησης και χρέωσης των χρηστών, καθώς και υποστήριξης της διαχείρισης διαπομπών. Η ρυθμιστική αρχή ενισχύει την εμπιστοσύνη των χρηστών στο σύστημα παρακολουθώντας τη συμπεριφορά των δικτυακών παρόχων και επεμβαίνοντας όταν κρίνεται απαραίτητο. Στην προτεινόμενη προσέγγιση, οι αποφάσεις εκτέλεσης διαπομπής ή κατανομής ροών κίνησης ενεργοποιούνται από πράκτορες, εκπροσώπους των χρηστών, που εκτελούνται στα κινητά τερματικά ή στο σταθερό δίκτυο, ανάλογα με την προέλευση των γεγονότων ενεργοποίησης. Η εκτέλεση των μηχανισμών απόφασης ανατίθεται σε πράκτορες που εκτελούνται στο σταθερό δίκτυο, για εξοικονόμηση των συνήθως περιορισμένων ενεργειακών και υπολογιστικών πόρων των τερματικών. Οι συγκεκριμένοι πράκτορες έχουν δυνατότητα μετακίνησης και εκτέλεσης σε πλατφόρμες που βρίσκονται πιο 'κοντά' στα τερματικά, από άποψης δικτυακής καθυστέρησης. Με τον τρόπο αυτό εξασφαλίζεται γρήγορη προσαρμογή του τερματικού στα διάφορα γεγονότα που απαιτούν την εκτέλεση διαπομπών σε επίπεδο δικτυακής διεπαφής ή ροής κίνησης. Στα πλαίσια της διατριβής προτείνεται η συνδυασμένη επιλογή υπηρεσιών μεταφοράς δεδομένων που προσφέρονται τόσο από δίκτυα πρόσβασης όσο και από δίκτυα κορμού. Για το σκοπό αυτό προτείνεται ένα μοντέλο δεδομένων για την περιγραφή των χαρακτηριστικών των υπηρεσιών, καθώς και μια διαδικασία για την αξιοποίησή τους από αλγορίθμους TFAP ή επιλογής δικτύου.

Η αξιολόγηση της επίδοσης της προτεινόμενης αρχιτεκτονικής έχει πραγματοποιηθεί μέσω ενός συστήματος προσομοίωσης που έχει υλοποιηθεί σε Java με χρήση του JADE (Java Agent Development framework) πλαισίου για την ανάπτυξη εφαρμογών βασισμένων σε πράκτορες λογισμικού. Η προσομοίωση έχει δυο στόχους: (α) εκτίμηση της καθυστέρησης που εισάγεται στην ενεργοποίηση της διαπομπής λόγω της επικοινωνίας και συνεργασίας των πρακτόρων του συστήματος και (β) μελέτη των επιπτώσεων της κινητικότητας των πρακτόρων στην απόδοση του συστήματος. Τα αποτελέσματα της προσομοίωσης δείχνουν μια περιορισμένη επιβάρυνση, της τάξης των 50ms, στο χρόνο ανίχνευσης της διαπομπής, η οποία δεν επιδρά σημαντικά στην απόκριση του τερματικού σε γεγονότα ενεργοποίησης διαπομπής. Η μελέτη της κινητικότητας των πρακτόρων, για διάφορους ρυθμούς κινουμένων πρακτόρων μεταξύ δυο πλατφορμών, ανέδειξε το μέγεθος της κατάστασης δεδομένων του πράκτορα ως ένα σημαντικό παράγοντα επίδοσης (μαζί με το ρυθμό μετακινούμενων πρακτόρων). Συγκεκριμένα, όσο μεγαλύτερο το μέγεθος της κατάστασης του πράκτορα, τόσο μεγαλύτερος χρόνος απαιτείται για τη μεταφορά και αποκατάσταση της κατάστασης εκτέλεσής του στη νέα πλατφόρμα (καθυστέρηση μετακίνησης). Βάσει των αποτελεσμάτων της προσομοίωσης, μέγεθος 4KB ή μικρότερο επιτρέπει υψηλούς ρυθμούς μετακινούμενων πρακτόρων (120 πράκτορες ανά δευτερόλεπτο) ενώ εξασφαλίζει μέση καθυστέρηση μετακίνησης κάτω από ένα δευτερόλεπτο. Θα πρέπει να σημειωθεί ότι η μετακίνηση του πράκτορα εκπροσώπου του χρήστη δεν επηρεάζει την εκτέλεση της διαπομπής καθώς πραγματοποιείται παράλληλα με αυτή. Επίσης, δεδομένου ότι η αποκατάσταση της εκτέλεσης του πράκτορα στη νέα πλατφόρμα απαιτείται για ενεργοποίηση νέων αποφάσεων εκτέλεσης διαπομπής, η καθυστέρηση μετακίνησης της τάξης του ενός δευτερολέπτου δεν επηρεάζει την απόκριση του τερματικού σε νέα γεγονότα ενεργοποίησης διαπομπής.

Ανοικτά ζητήματα που προέκυψαν από τη διατριβή για περαιτέρω διερεύνηση είναι: (α) η εισαγωγή ενός προσεγγιστικού αλγορίθμου με εγγυημένο σφάλμα προσέγγισης για το πρόβλημα TFAP, (β) η επέκταση της διατύπωσης του προβλήματος για την υποστήριξη ροών με εναλλακτικές απαιτήσεις σε χωρητικότητα, καθώς και ροών που μπορούν να διασπαστούν σε περισσότερες της μιας δικτυακές διεπαφές, (γ) εισαγωγή ενός αλγορίθμου καθορισμού του άνω ορίου στην κατανάλωση ενέργειας ενός τερματικού δεδομένης της τρέχουσας κατάστασης, του περιβάλλοντός του και του προφίλ κίνησης του χρήστη.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου Αναπληρωτή Καθηγητή Μανόλη Γιακουμάκη για την επιστημονική του καθοδήγηση κατά τη διάρκεια εκπόνησης της διατριβής και την ηθική του υποστήριξη σε δύσκολες στιγμές της πορείας ανάπτυξης αυτής της δουλειάς. Επιπλέον, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή Νίκο Μαλεύρη και τον Επίκουρο Καθηγητή Γιώργο Ξυλωμένο, μέλη της τριμελούς επιτροπής μου, για τις πολύ χρήσιμες υποδείξεις τους σχετικά με την βελτίωση της οργάνωσης και παρουσίασης της διατριβής. Ειδικές ευχαριστίες οφείλω στον Επίκουρο Καθηγητή Σταύρο Τουμπή για την συνδρομή του σε μια κρίσιμη φάση ανάπτυξης της δουλειάς που περιλαμβάνεται στο Κεφάλαιο 3 της διατριβής. Τα ενθαρρυντικά σχόλια και οι υποδείξεις του σχετικά με την ενίσχυση του κινήτρου αυτής της δουλειάς, έπαιξαν καθοριστικό ρόλο στην τελική δημοσίευση των αποτελεσμάτων της. Επίσης, σημαντική ήταν και η συμβολή του Λέκτορα Βαγγέλη Μαρκάκη σε ζητήματα σχετικά με την ορθότητα της απόδειξης για την πολυπλοκότητα επίλυσης του προβλήματος κατανομής ροών κίνησης. Τέλος, θα ήθελα να ευχαριστήσω το Γιώργο Λουκαρέλλι και τον Νίκο Διαμαντίδη για τις πολύ χρήσιμες συζητήσεις που είχαμε, σε διάφορες χρονικές στιγμές, πάνω σε ζητήματα που αφορούν τη διατριβή.

Σημαντικό ρόλο στην ολοκλήρωση αυτής της δουλειάς είχαν οι φίλοι μου και συνάδελφοι, Δημήτρης Βάσσης και Χρήστος Δουλκερίδης, τους οποίους ευχαριστώ για τη συνεργασία και τις πολύ εποικοδομητικές συζητήσεις που είχαμε αρκετά συχνά.

Τέλος, ευχαριστώ την οικογένειά μου για την υποστήριξή της σε όλη τη χρονική διάρκεια εκπόνησης της διατριβής.

Abstract

Multi-radio mobile communication devices are increasingly gaining market share due to the diversity of currently deployed and continuously emerging radio access technologies. Multi-homing support in multi-radio terminals, i.e., simultaneous use of two or more radio interfaces, provides improved user experience through increased bandwidth capacity availability and reliability of wireless access. Furthermore, optimized assignment of application traffic flows to available interfaces and radio access bearer services contributes to economic and power consumption efficiency. The thesis studies the traffic flow assignment problem (TFAP) in a mobile node, multi-homed through a set of different technology radio interfaces. It introduces an analytical formulation for the problem and proves its hardness through reduction from the Multiple Knapsack Problem with Assignment Restrictions. Problem solutions are approximated with a heuristic algorithm that is based on local search and is characterized by efficient execution times for a wide set of realistic problem sizes. The quality of approximation is rather satisfactory and is evaluated through comparison of heuristic and exact solutions for a large set of randomly generated problem instances. Moreover, an evaluation of the approach through simulation supports these findings and provides an estimation of the associated mobility management overhead that is limited and allows real deployment of the decision mechanism.

The employment of advanced network selection (or handover decision) or TFAP algorithms cannot be based solely on an end-host infrastructure. The decision mechanisms require both locally available information (e.g. application traffic requirements, user preferences etc.) and location-based network information that is not practical to be retrieved by the mobile terminal exclusively through active scanning. The reason is that active scanning is time consuming and inefficient in terms of energy consumption. Moreover, reliable and in-time information on resource availability of available access networks may not be provided by a single network operator, e.g., the home operator of a mobile user, as it has no incentives to provide it and consequently let its customers utilize third-party services. In this thesis a system architecture is also proposed for supporting the execution of handover decisions or TFAP algorithms. The architecture spans multiple administrative domains and is based on software agents. The software agents represent the users, the network operators, a Multi-Access Provider (MAP) and the regulatory authority. In the proposed approach, handover or traffic flow assignment decisions are delegated to software agents that are user representatives. Decision making is initiated by user agents that execute either in the terminal or the network, depending on the source of handover triggering events. On the other hand, execution of decision algorithms takes place in the network for saving terminal's usually limited power and computational resources. Performance evaluation of the architecture has been performed through a simulation system with focus on the impact on handover latency. The results are promising for the feasibility of the proposed architecture.

Contents

List of Tables

List of Figures

Acronyms

AAA Authentication, Authorization and Accounting.

AB Access Bearer.

ABC Always Best Connected.

ABS Always Best Served.

ACL Agent Communication Language.

AF-agent Access Facilitator agent.

AKA Authentication and Key Agreement.

AMS Agent Management System.

ANDSF Access Network Discovery and Selection Function.

AP Access Point.

API Application Programming Interface.

AR Access Router.

ASN Access Service Network.

BAG Bandwidth Aggregation.

BER Bit Error Rate.

BID Binding Identification Number.

BIP Binary Integer Programming.

CB Core Bearer.

CIM Common Information Model.

CM-agent Connection Manager agent.

CMT Concurrent Multi-path Transfer.

CN Correspondent Node.

CoA Care-of Address.

CSN Connectivity Service Network.

CWN Composite Wireless Network.

DF Directory Facilitator.

DMTF Distributed Management Task Force.

DSMIPv6 Dual Stack MIPv6.

E-UTRAN Evolved-UTRAN.

EAP Extensible Authentication Protocol.

EPC Evolved Packet Core.

ePDG evolved Packet Data Gateway.

EPS Evolved Packet System.

FA Foreign Agent.

FID Flow Binding Identification.

FIPA Foundation for Intelligent Physical Agents.

GAN Generic Access Network.

GAS Generic Advertisement Service.

GERAN GPRS/EDGE Terrestrial Radio Access Network.

GGSN Gateway GPRS Support Node.

GIS Geographical Information System.

GPS Global Positioning System.

GTP GPRS Tunneling Protocol.

HA Home Agent.

HNP home network prefix.

HoA Home Address.

HPLMN home PLMN.

HRPD High Rate Packet Data.

HSGW High Speed Gateway.

HSS Home Subscriber Subsystem.

HTTP Hypertext Transfer Protocol.

I-WLAN Interworking-WLAN.

IETF Internet Engineering Task Force.

IIOP Internet Inter-Orb Protocol.

ILP Integer Linear Programming.

IMSI International Mobile Subscriber Identity.

IWK Interworking.

JADE Java Agent Development framework.

JVM Java Virtual Machine.

KIF Knowledge Interchange Format.

KQML Knowledge Query and Manipulation Language.

LMA Local Mobility Anchor.

LTE Long-term Evolution.

MAG Mobility Access Gateway.

MAHO Mobile Assisted Handover.

MAP Multi-Access Provider.

MCHO Mobile Controlled Handover.

MEXT WG Mobility EXtensions for IPv6 Working Group.

MICS Media Independent Command Service.

MIES Media Independent Event Service.

MIH Media Independent Handover.

MIH PoS MIH Point of Service.

MIH_SAP media independent handover service access point.

MIHF MIH Function.

MIIS Media Independent Information Service.

MIPv6 Mobile IPv6.

MKAR Multiple Knapsack problem with Assignment Restrictions.

MMT Mobile Multi-mode Terminal.

MN Mobile Node.

MO management object.

MONAMI6 WG Mobile Nodes and Multiple Interfaces in IPv6 Working Group.

MOOP Multi-Objective Optimization Problem.

MT Mobile Terminal.

MTS Message Transport Service.

NCHO Network Controlled Handover.

NM-agent Network Monitor agent.

NP-agent Network Provider agent.

NRM Network Reconfiguration Manager.

OMA DM Open Mobile Alliance Device Management.

OSM Operator Spectrum Manager.

P-agent Profile agent.

P-GW Packet Gateway.

PDN Packet Data Network.

PLMN Public Land Mobile Network.

PMIPv6 Proxy Mobile IPv6.

PoA Point of Access.

PS Packet Switched.

QoS Quality-of-Service.

RAN Radio Access Network.

RAT Radio Access Technology.

RDF Resource Description Framework.

RMC RAN Measurement Collector.

RRC RAN Reconfiguration Controller.

RRM Radio Resource Management.

RSS Received Signal Strength.

RTT Round Trip Time.

S-GW Serving Gateway.

SAP service access point.

SCTP Stream Control Transmission Protocol.

SGSN Serving GPRS Support Node.

SIM Subscriber Identity Module.

SLA Service-Level Agreement.

SSID Service set identifier.

SSPN Subscription Service Provider Network.

TFAP Traffic Flow Assignment Problem.

TMC Terminal Measurement Collector.

TRC Terminal Reconfiguration Controller.

TRM Terminal Reconfiguration Manager.

UICC Universal Integrated Circuit Card.

UMA Unlicensed Mobile Access.

UML Unified Modeling Language.

USIM Universal Subscriber Identity Module.

UTRAN UMTS Terrestrial Radio Access Network.

VPLMN visited PLMN.

W-APN Wireless Access Point Name.

WAP Wireless Application Protocol.

Chapter 1

Introduction

Mobile Internet access is enabled through a variety of Radio Access Technologies (RATs), e.g., UMTS, LTE, IEEE 802.16m/e, IEEE 802.11 etc., that are characterized by diversity in service attributes (e.g., QoS provision, peak data rate, capacity), service range (local area, metropolitan or wide area) and deployment costs. The term RAT or *radio interface* is defined in [IEE09a] as specifications of an air interface that shall be fulfilled in order to setup and maintain connection between terminal and base station and may be characterized by multiple access method, modulation etc. In this thesis the term *radio interface* is used to refer to the user terminal equipment that enables communication with base stations and may support one or more RATs. The base stations that support a certain RAT along with the network that connects them to the packet-based core network or external networks is termed as Radio Access Network (RAN) [IEE09a].

Usually the various RATs act competitively to each other, e.g., WLANs act as a cheap and high speed alternative to GPRS for data traffic. However, the traffic requirements in terms of volume and QoS, following the increasing market penetration of mobile communication devices with advanced processing and multimedia capabilities, shift the focus towards the opposite direction, i.e., combined and complementary use of RANs, corresponding to different RATs, for capacity increase and improved user experience. From an end-user perspective, complementary use of different RAN types involves: (1) access to their services through a single subscription and billing account, (2) transparent utilization of the most efficient available access network(s) on the basis of preferences related to economic cost and application performance, (3) seamless mobility across them, if needed, for coverage reasons or due to enforcement of user preferences. Network operators seek maximization of their return on investment through efficient allocation of available radio access resources. From a network operator perspective, complementary use of different RANs involves their load balancing for the maximization of the number of admitted users without degradation of the perceived QoS of ongoing user sessions. Load balancing involves: (i) admission of newly arriving user sessions to an appropriate RAN, (ii) seamless redirection of already served traffic sessions across available RANs.

The combined use of heterogeneous RANs requires the integration of functions such as Authentication, Authorization and Accounting (AAA), mobility management, resource management etc., of different systems. Such integration is referred to as interworking and is a general trend in the evolution of the specifications of major mobile communications' systems' architectures. Starting from 3GPP Release 7 and WiMAX Release 1.0, network architecture specifications provide for interworking with third-party radio access networks, especially WLANs (IEEE802.11). Their primary focus is on seamless mobility across different technology RANs.

Seamless inter-RAT mobility is a basic requirement towards 4G [BCG09], where 4G represents a capability of wireless access that is constantly optimized given the availability of RANs, application traffic requirements and user preferences. Optimization refers to the utilization of the most appropriate radio interface and wireless access service for serving a Mobile Terminal (MT)'s application traffic requirements. A more advanced capability is represented by the Always Best Connected (ABC) concept [BCG09, GJ03] where multiple radio interfaces may be activated for supporting a MT's optimal connectivity state. Thus, ABC requires multi-homing support, in addition to seamless inter-RAT mobility, a feature that currently cannot be combined with node mobility as it is not supported by Mobile IPv6 (MIPv6) and other mobility management protocols. The Internet Engineering Task Force (IETF) MONAMI6 WG has identified the benefits that mobile host multi-homing offers to both end users and network operators [EMWK08] and its successor, IETF MEXT WG, is working towards enhancing MIPv6 with multi-homing support.

An ABC-enabled Mobile Multi-mode Terminal (MMT) extends its degrees of freedom related to the adaptation of its connectivity state to the changing traffic requirements and wireless networking context. For instance, the range of options for responding to the arrival of a traffic flow, when spare capacity in active radio interfaces is not available, may include: (a) activation of an inactive radio interface and its attachment to an appropriate radio bearer service, (b) horizontal handover on an active radio interface towards a higher capacity bearer service, (c) redirection of one or more traffic flows, already served by one interface, to another interface for best utilization of available bandwidth capacity etc. The set of available options on each occasion depends on the wireless context, the MMT's traffic load and hardware configuration. Moreover, each alternative may have different impact on the fulfillment of user preferences and especially on economic efficiency and energy autonomy. Thus, evaluation and determination of the optimal operational state requires advanced and fast executing decision algorithms. Execution efficiency is required due to the frequently occurring triggers for decision making that include changes in served traffic, network conditions and device status (e.g., battery lifetime).

Assume a MMT that is equipped with different technology radio interfaces, e.g., 3GPP, WLAN, WiMAX. The MMT has either the role of (a) an end-host that serves the traffic generated by user applications, or (b) a mobile router that acts as an Internet gateway in a Local Area or Personal Area Network. The MMT operates in an area served by multiple RANs corresponding to different RATs and is capable of connecting to anyone of them. Despite the fact that current mobile handsets and notebooks usually combine two or three radio interfaces, advanced multi-interface devices for business communications purposes are starting to emerge. Figure 1.1 shows an example of a "bandwidth bonding" appliance (Portabella 2242 and 141), created by Mushroom Networks, Inc. [Mus09], that provides high capacity wireless access to business users by aggregating the offered bandwidth of multiple cellular connections. This thesis studies in chapter 3 the problem of assignment of application traffic flows (either inbound or outbound) of a MMT to appropriate radio interfaces and radio bearer services in a way that: (i) satisfies the traffic flows' QoS requirements and the bearer services' capacity constraints, and (ii) establishes the best trade-off between economic cost and power consumption. The problem will be henceforth referred to as Traffic Flow Assignment Problem (TFAP). The economic cost factor of TFAP corresponds to network usage cost, while power consumption is due to the operation of active radio interfaces. Due to the dynamic nature of problem parameters, the MMT faces iteratively TFAP instances of variable size during its operation lifetime. The thesis provides an analytical formulation of TFAP and a study on its complexity through reduction to the Multiple Knapsack Problem with Assignment Restrictions (MKAR). Since MKAR is NP-hard, TFAP is also NP-hard and approximation algorithms are required for fast derivation of problem solutions. Moreover, a heuristic local search algorithm is introduced that is characterized by efficient execution times for a wide set of realistic problem sizes.

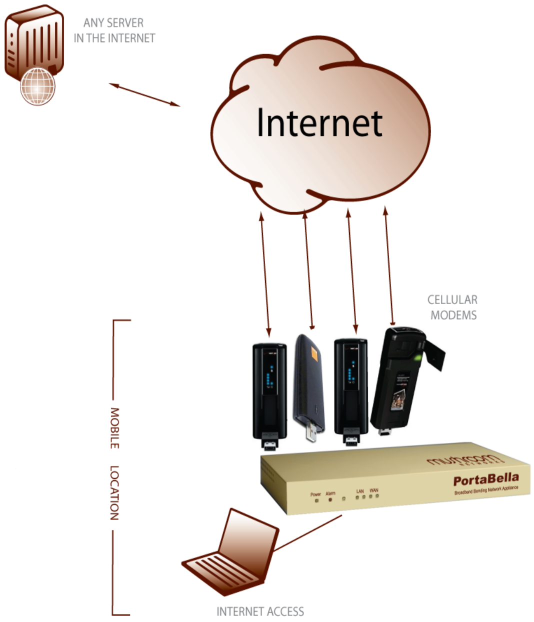

Figure 1.1: PortaBella BBNA2242 (Broadband Bonding Network Appliance). *Source: http://www.mushroomnetworks.com*

Network selection (or handover decision) represents a basic part of the handover procedure that ensures service continuity as the MMT roams across service areas (cells) of the same or different radio access systems. A handover across RANs of different RAT is termed vertical handover, as opposed to a horizontal handover that takes place among points of access of the same technology and provider. The availability of different radio access overlays, in a heterogeneous network setting, broadens the scope of network selection from preserving connection quality to serving user objectives such as economic efficiency, energy autonomy etc. Network selection represents a special case of TFAP that applies in single-homed MMTs, i.e, in MMTs without capability of simultaneous use of multiple radio interfaces. While handover decision is followed by handover execution, the enforcement of TFAP decisions may involve the execution of one or more horizontal, vertical or flow handovers, depending on the number of available and activated radio interfaces in the MMT.

The deployment of TFAP or advanced network selection algorithms require an execution context that provides: (a) timely delivery of notifications that trigger the decision mechanism, (b) sufficient processing power resources for fast algorithm execution and reaction to events, (c) access to all information required for algorithm execution. The latter comes from multiple sources and administrative domains, while the events that trigger problem solving span multiple layers of the protocol stack. Thus, a distributed application layer infrastructure is required, as it is also proposed in [GJ03] where the requirements for an ABC service are set. Moreover, a trustworthy implementation of this capability cannot be offered by a single network provider. The reason is that a network provider has no incentives to provide reliable and in-time information regarding available wireless networks and consequently let its customers utilize third-party services. A viable solution should: (a) incorporate various wireless operators, (b) support market competition through easy integration of new entrants, (c) adopt a common, unambiguous information schema for interoperability of the exchanged information (e.g., descriptions of network capabilities, so as to enable effective decision making), (d) build on a commonly accepted model of trust relationships so as to be relied upon by users and network operators. In chapter 4 of this thesis a system architecture is proposed that takes into account these requirements. The architecture spans multiple administrative domains and is based on software agents. The software agents represent the users, the network operators, a Multi-Access Provider (MAP) and the regulatory authority. MAP is a business entity that maintains roaming agreements with network operators and enables user utilization of their services through a single subscription. Moreover, MAP serves AAA and billing purposes and supports inter-domain mobility management. The regulator enhances user trust by monitoring the behavior of the operators and intervening when required. In the proposed approach, handover or traffic flow assignment decisions (in single-homed or multi-homed hosts respectively), are delegated to software agents that are user representatives. Decision making is initiated by user agents that execute either in the terminal or the network side, depending on the source of handover triggering events. On the other hand, execution of decision algorithms takes place in the network for saving a terminal's usually limited power and computational resources.

1.1 Base assumptions and technological context

This thesis focuses on decision making support for optimized traffic flow assignment in ABC-enabled mobile hosts equipped with two or more different technology radio interfaces. An

ABC-enabled mobile host may have the role of an end-host or a mobile router and is assumed to incorporate the following capabilities: (a) support of seamless mobility across RANs of different RAT, (b) multi-homing support and (c) fine-grained mobility at a traffic flow level in cases that two or more radio interfaces are simultaneously activated. Sections 1.1.1 and 1.1.2 summarize the state of the art in architecture and protocol specifications of mobile communications systems that are relevant to the realization of these assumptions.

1.1.1 Towards seamless inter-RAT mobility

3GPP identifies six interworking scenarios, each one representing an incremental degree of WLAN integration in the 3GPP service offering [3GP09b]. Each scenario identifies service and operational capabilities required for each degree of interworking:

- Scenario 1: Common Billing and Customer Care
- Scenario 2: 3GPP system based Access Control and Charging
- Scenario 3: Access to 3GPP system's Packet Switched services
- Scenario 4: Service continuity
- Scenario 5: Seamless service continuity
- Scenario 6: Access to 3GPP system's Circuit Switched Services

As no use cases have been identified for Scenario 6, it is not considered by 3GPP for further development. Thus, the highest degree of interworking requirements (according to 3GPP) is represented by Scenario 5 where seamless handover is enabled between two different technology radio access networks. Inter-RAT handover is also called *vertical handover* due to service mobility between different technology radio access overlays available in a certain MMT location [SK98]. Seamless vertical handover is characterized by minimal service disruption and is also known as make-before-break or *soft handover*. A *hard handover* (or break-before-make handover), on the other hand, is characterized by interruption of UE's connections for a short time period (usually 1 to 10 seconds) during handover execution. The higher the degree of interworking between the source and the target mobile communications systems, the more seamless is the vertical handover among them.

Soft handover realization depends on the level of interworking of the source and target systems. Thus, soft handovers can be further categorized into: (a) *single radio* handovers and (b) *dual radio* handovers, on the basis of the number of active radio interfaces involved in handover execution. A single radio handover involves registration and reservation of the required resources in the target radio access network through the radio interface that is connected to the source RAN. Once the target RAN is prepared to admit the MMT's connection, the radio interface corresponding to the target RAN is activated while the other is switched off. This requires tight interworking between the source system's core network with radio access network elements of the target system. For this reason, the IEEE 802.21 WG is working towards generic and media independent protocols for single-radio handover realization [IEE10a]. A dual radio handover requires looser coupling between source and target systems. In dual radio handovers the source radio interface serves data traffic while the target radio interface registers to the target RAN. Once registration is complete and data traffic starts to flow through the target radio interface, the source radio interface is deactivated. Figure 1.2 depicts the aforementioned handover types and the interworking levels at which they are enabled.

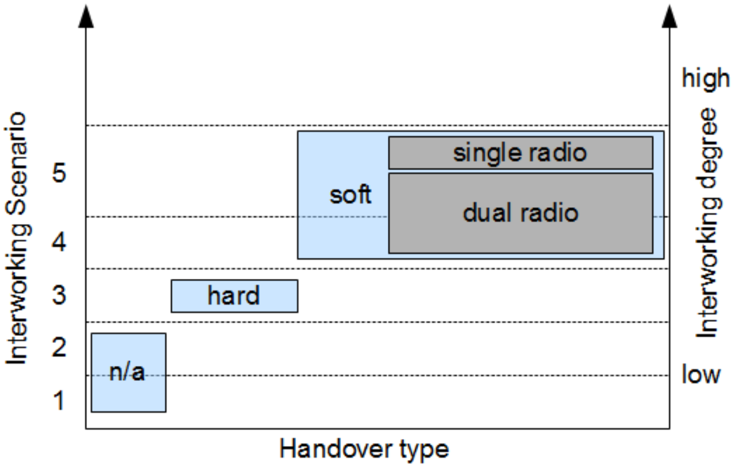

Figure 1.2: Handover Categorization

The interworking requirements represented by each scenario are used by 3GPP as a reference for the characterization of the various interworking solutions that are proposed in UMTS and LTE specifications and can be applied for other radio access technologies as well (e.g., WiMAX, IEEE 802.11). Widely deployed wireless access technologies (3GPP, WiMAX, WiFi) gradually incorporate in their specifications interworking capabilities with third-party RATs and their evolution heads towards scenario 5 interworking.

3GPP specifications for UMTS (Release 7) introduce an architecture for interworking between the GPRS core network and a WLAN access system. The architecture, known as Interworking-WLAN (I-WLAN) architecture, specifies two interworking configurations: (a) *WLAN Direct IP access* for access control and charging through a 3GPP system of the services provided by a WLAN and (b) *WLAN 3GPP IP access* for enabling access to 3GPP Packet Switched (PS) services through a WLAN access network [3GP08a]. WLAN Direct IP access and WLAN 3GPP IP access support the interworking requirements represented by Scenarios 2 and 3 respectively. Scenario 4 requirements for service continuity during vertical 3GPP-WLAN handovers in 3GPP IP access were later introduced in Release 8 specifications [3GP09d]. Specifically, in the Release 8 I-WLAN architecture, a Home Agent (HA) network element is introduced in the GPRS core network for handling mobility between 3GPP and WLAN access networks. Mobility is enabled with the Dual Stack MIPv6 (DSMIPv6) protocol in a transparent manner. Finally, Scenario 6, tight interworking between 3GPP and IP access networks, is enabled with the Generic Access Network (GAN) specification, also known as Unlicensed Mobile Access (UMA) [3GP09c]. The specification enables any generic IP access network to interwork with the 3GPP core network as an ordinary UMTS Terrestrial Radio Access Network (UTRAN) or GPRS/EDGE Terrestrial Radio Access Network (GERAN) access network.

Release 8 of 3GPP specifications introduced enhancements to the UMTS radio access and core network that represent an evolution of the system known as Long-term Evolution (LTE) or Evolved Packet System (EPS). The evolved radio access network is named Evolved-UTRAN (E-UTRAN) while the new all-IP core network architecture is called Evolved Packet Core (EPC). A basic LTE requirement was its interworking with non-3GPP access networks. Thus, access to LTE Packet Switched (PS) services and mobility between E-UTRAN and non-3GPP access networks was a basic design objective of the evolved system. Non-3GPP access networks are categorized in the LTE specifications into trusted and un-trusted [3GP09a]. Trusted non-3GPP access networks are characterized by their capability of performing 3GPP defined authentication, while un-trusted ones perform authentication through a secure tunnel that is established between the MMT and an EPC interworking gateway, evolved Packet Data Gateway (ePDG).

The interworking solutions for trusted and un-trusted non-3GPP networks cover Scenario 4 interworking requirements and are also called Handovers without Optimizations.

Especially for specific trusted non-3GPP access networks Scenario 5 interworking is supported through tighter interworking solutions that are specified under the name Handovers with Optimizations [3GP10b]. A differentiating factor between Handovers with Optimizations and Handovers without Optimizations is the MMT pre-registration feature that is offered by the first set of solutions. Pre-registration enables the MMT to register and allocate resources in the target RAN without establishing a connection to it but through forwarding the required signaling traffic via the source system's core network. For the time being cdma2000® High Rate Packet Data (HRPD)[1]is the only system for which such tight integration is supported. Work is also under progress for tight integration of the WiMAX access network to the LTE core.

With regard to WiMAX, the WiMAX Forum has included in Release 1.0 WiMAX specifications its interworking with 3GPP systems. The WiMAX-3GPP interworking architecture is based on the I-WLAN architecture as it is specified by 3GPP in [3GP08a]. In [WiM08] Direct IP and 3GPP IP access are specified through a WiMAX access network[2]. The degree of WiMAX-3GPP interworking represented by these solutions correspond to Scenarios 2 and 3 respectively. More advanced scenarios of interworking are out of scope of WiMAX Release 1.0 interworking specifications.

The IEEE 802.11 Working Group is also working on specifications for 802.11 Access Point (AP) interworking with external, possibly different technology, networks that are referred to as Subscription Service Provider Networks (SSPNs). The Technical Group u (TGu) is responsible for the interworking specifications that are prepared as an amendment to the base IEEE802.11 standard (IEEE 802.11u) [IEE10b]. The amendment provides for:

- network discovery and selection, i.e., discovery by a MMT of network infrastructure capabilities accessible through available APs without associating with them and selection on the basis of this information of the most appropriate for serving its needs. For instance, a MMT may discover APs that provide 3GPP IP access through a specific 3G operator.
- QoS mapping of layer-3 service levels of other SSPNs to IEEE 802.11 wireless access service levels.
- emergency services,

---

[1] High Rate Packet Data

[2] ASN - Access Service Network

- service interface between the AP and the SSPN (transfer of user policies that affect authentication, authorization, admission control, transfer of instructions on service provision, e.g., termination of a connection to a MMT).

The amendment specifies the loose coupling of 802.11 APs with external networks and standardizes WLAN specific parts of the 3GPP I-WLAN specification such as network selection and discovery and transfer of authentication and authorization policies to the WLAN.

Table 1.1 provides a summary of the state-of-the art in the standardization of interworking of widely deployed wireless access technologies.

1.1.2 Multi-homed multi-radio wireless access

IETF has recently documented the benefits of simultaneous use of multiple interfaces and global addresses in mobile nodes [EMWK08]. The capability of simultaneous use of two or more global IP addresses by a mobile node is known as multi-homing support. Multi-homing can be enabled either by the use of multiple physical network interfaces each one configured with a different IP address or through the configuration of multiple IP addresses to a single physical network interface due to the advertisement of different network prefixes in its current link. The benefits of multi-homing can be reached once a multi-homed node incorporates a set of capabilities, namely: (a) reliability, (b) load sharing and (c) flow distribution [MWE+08].

Reliability denotes the capability of using multiple radio interfaces connected to different access networks for: (a) connection recovery in case that a currently used access network becomes unavailable, or (b) n-casting (usually bi-casting) of traffic flows belonging to critical applications over different RANs, especially in cases that provided QoS is inferior to that required by applications (e.g., in locations with poor coverage). Load sharing refers to the distribution of traffic load (originating from or terminating to a certain Mobile Node (MN)) over multiple RANs, through respective radio interfaces, for load-balancing reasons or for access to increased capacity. Finally, flow distribution refers to the capability of selective redirection of traffic flows across different radio interfaces for reasons of load balancing or due to user or operator defined policies. As MIPv6 (and Proxy Mobile IPv6) do not natively enable all these capabilities [MWE+08], IETF working groups that are involved in their specification are currently working on relevant extensions.

Table 1.1: Standardization activities in RATs interworking.

Interworking configuration (base / IWK system)	**Mobility Management Protocol**	**Local Mobility Anchor (Main anchor point)**	**Mobile Access Gateway (Local anchor point)**	**Handover type**	**Network selection and discovery mechanism**	**3GPP IWK Scenario**
LTE / UTRAN, GERAN	GTP	P-GW (GGSN)	S-GW	optimized, single radio, network controlled	MT measurements, based on signal quality and RRM decision	5
3GPP Rel.8 / WLAN (I-WLAN)	DSMIPv6	HA	WLAN AR	make before break, mobile controlled, non optimized, dual radio	automatic or manual network selection, based on list of available I-WLANs and PLMNs, 802.11u and GAS, 802.1x and EAP	4
LTE / Trusted non-3GPP w/o Optimizations	PMIPv6 or DSMIPv6, MIPv6	P-GW	MAG provided by the trusted non-3GPP RAN	make before break, mobile controlled, non optimized, dual radio	ANDSF	4
LTE / Un-trusted non-3GPP	PMIPv6 or DSMIPv6, MIPv6	P-GW	ePDG	same as above	ANDSF	4
LTE / Trusted non-3GPP with optimizations (cdma2000 HRPD)	PMIPv6	P-GW	S-GW or HSGW	make before break,single-radio, optimized (pre-registration)	ANDSF	5
WiMAX Release 1.0 / 3GPP Rel. 7	PMIPv6, MIPv6	HA in CSN	FA in ASN	n/a	n/a	2, 3
3GPP Rel. 7 / WLAN (WLAN Direct IP access)	n/a	n/a	n/a	n/a	n/a	2
3GPP Rel. 7 / WLAN (WLAN 3GPP IP access)	n/a	n/a	n/a	n/a	n/a	3
3GPP Rel. 8 / Generic IP access networks (GAN)	GTP	GGSN	SGSN	Seamless	n/a	6
802.11u	n/a	n/a	n/a	n/a	n/a	2

RFC 5648 [WDT$^+$09], that is currently a proposed IETF standard, contributes towards this direction by enabling simultaneous registration of multiple Care-of addresses (CoAs) to a single Home Address (HoA). Thus, a MN may register with its HA, as care-of addresses, all the different global IP addresses that are configured to its active radio interfaces. The specification introduces the Binding Identification Number (BID) concept that identifies the different bindings of CoAs to the MN's HoA. Usually, each BID represents an activated radio interface that the user requires to be reachable through a specific HoA. A MN may register alternative bindings to the HA or to a Correspondent Node (CN) and either one of the available bindings can be used for communicating with the MN. Moreover, a MN may perform binding updates in case that the CoA corresponding to a BID has changed (e.g., due to a handover on the respective radio interface).

RFC 5648 does not specify algorithms or mechanisms that a CN or HA may utilize for assigning traffic flows to the available BIDs of a certain MN. A protocol for the definition and transfer of traffic flow assignment policies is the focus of the Flow Bindings Internet Draft [STM$^+$10] that complements RFC 5648 and is intended for standardization. This protocol enables the MN to register to the HA or to a CN its preferences related to the distribution of incoming traffic to its active radio interfaces. It is based on the Flow Binding Identification (FID) concept that defines a binding of a traffic flow to a BID (or set of BIDs in case that n-casting is required). A MN may introduce multiple FIDs each one specifying the handling of a different traffic flow. Moreover, a MN may update the BIDs associated with a FID enabling thus mobility at a traffic flow granularity (*flow handover*). Note that the term traffic flow refers to a set of IP packets that match a specific traffic selector [STM$^+$10]. A traffic selector describes packet attributes and their values that are shared among the flow's packets such as source and destination IP addresses, source and destination ports, transport protocol number and other IP or higher layer header fields. The format of traffic selectors is currently specified in [TGSM10].

Proxy Mobile IPv6 (PMIPv6), a network-based solution standardized by IETF for localized mobility management, has native support for MN multi-homing [GLD$^+$08]. A MN that is authorized to register to a PMIPv6 mobility domain acquires its address(es), through the network's address configuration mechanism, from a set of home network prefixes (HNPs) that are uniquely assigned to it. A different set of HNPs is (statically or dynamically) assigned to each network interface of a MN that connects to a PMIPv6 domain, in analogy to the assignment of HoAs in standard MIPv6. PMIPv6 handles host mobility in a transparent manner, i.e., without involvement of the MN in IP mobility management signaling. The protocol is based on two basic functional elements: (a) the Local Mobility Anchor (LMA) that has the role of the HA in a PMIPv6 domain by maintaining MN reachability state and acting as the topological anchor point for the MN's HNPs, (b) the Mobility Access Gateway (MAG) that is deployed in each access router and handles mobility management signaling on the behalf of the MN. Specifically, the MAG updates the MN's reachability (binding) state to the LMA by tracking its movements and emulates the MN's home link by providing solicited Router Advertisements with the same address configuration properties (e.g., HNPs, default router) across the entire PMIPv6 mobility domain. A MN that moves across RANs served by different MAGs does not notice any change in its Layer 3 connectivity status as the new MAG replies to MN's Router Solicitations with the same address configuration properties that are retrieved from the LMA. The lookup key that is used to locate a MN's binding state in the LMA is a Mobile Node Identifier (e.g., a Network Address Identifier or a MAC address) that becomes available to a MAG either by the MN during authentication or the previous MAG. A MN-Identifier and the HNP assigned to a MN characterize a mobility session in PMIPv6.

PMIPv6 natively supports mobile node multi-homing through a single physical interface (radio interface) by allowing the configuration of multiple IP addresses on the interface from the MN's set of HNPs. However, addresses that belong to the HNPs assigned to a certain mobility session may not be configured to more than one radio interfaces. Thus, a MN engaged in a single mobility session may not simultaneously utilize two or more radio interfaces, although handovers across different interfaces are supported. In this case the source radio interface is deactivated after handover execution. Assume a MN that simultaneously activates a second radio interface and registers to the same PMIPv6 domain. A new mobility session is, then, spawned, that is characterized by a different MN-Identifier (e.g., the MAC address of the newly activated radio interface), and a new set of HNPs is assigned to the MN. PMIPv6 handles independently for each mobility session the continuity of their respective flows as the MN's radio interfaces roam across RANs served by different MAGs. A vertical handover on one of its radio interfaces will redirect the flows of its respective mobility session to the second radio interface, resulting, thus, in a single active radio interface serving the flows that belong to two mobility sessions. With appropriate signaling support from the MN (that is out of scope of RFC 5213) the flows of either mobility session could be turned back to the first radio interface after a second vertical handover. Flow mobility is, thus, enabled at the granularity of a mobility session, i.e., it is not possible to move a traffic flow from one mobility session to the other, but instead the flows of each mobility session are redirected as a set across RANs of the same or different RAT. The NetExt working group that is working on extensions to PMIPv6 is currently evaluating enhancements to PMIPv6 that will allow fine-grained mobility of multi-homed MN's at a traffic flow level [MG10, BJK+10].

PMIPv6 has been adopted by 3GPP Release 8 EPS for network controlled mobility management across 3GPP access networks and non-3GPP trusted/untrusted access networks [3GP09a]. Although a MN may configure multiple HoAs to one or more Packet Data Networks (PDNs), it is not possible to route traffic from a single or multiple PDNs to more than one access systems simultaneously. Thus, mobile node multi-homing through multiple radio interfaces is not currently supported. The same holds in cases that DSMIPv6 [Sol09] is used for MN controlled mobility management across different radio access systems. Regarding the Release 8 I-WLAN interworking architecture, that is also based on DSMIPv6 for session continuity across 3GPP and WLAN access systems, simultaneous use of more than one radio interfaces is not possible. 3GPP has recently published a technical report [3GP09e] that studies solutions for the support of a) simultaneous use of two different access systems for serving the traffic of one or more PDN connections and b) IP flow mobility across the different access systems. The studied solutions are based on PMIPv6 and DSMIPv6 and are applicable to both I-WLAN and EPS interworking architectures. The report proposes as most appropriate the DSMIPv6 based solutions that incorporate MIPv6 extensions for this purpose [STM+10] [WDT+09]. The solution is further specified and considered for adoption in Release 10 LTE specifications [3GP10c].

1.2 Thesis Contribution

The contribution of this thesis includes the following:

- Unified handling of the subproblems of radio interface activation, network selection and traffic flow distribution in a mobile node with multiple radio interfaces and multi-homing capabilities, that operates in heterogeneous network setting. This thesis specifies and provides an analytical formulation of the resulting problem, named Traffic Flow Assignment Problem (TFAP), that involves the joint optimization of operational cost (network usage charges) and energy consumption of the mobile node. The formulation incorporates additional constraints, with reference to related approaches in the literature, that results in a more realistic modeling of the problem domain.

- Solution method for TFAP that involves its transformation to a single-objective optimization problem. The method prescribes the determination of an upper limit in the mobile node's power consumption, on the basis of device and user related factors, and the conversion of the power consumption objective to a problem constraint. The formulation of the single-objective problem is validated through its mapping to a binary integer programming representation and solution of sample problem instances in the Lingo environment [LIN09].

- Complexity analysis of TFAP through reduction from the Multiple Knapsack Problem with Assignment Restrictions that is NP-Hard.

- Specification of a heuristic algorithm, that is based on local search, for approximating TFAP problem solutions. The algorithm establishes a good trade-off between quality of results and performance of execution. It's computational efficiency allows its deployment to resource constrained mobile devices or to mobile routers that iteratively face considerably larger problem instances.

- Evaluation of the merits of optimized flow assignment in a simulated environment. Towards this purpose an event-driven simulator has been implemented and the heuristic algorithm is compared against an alternative that has been proposed in a similar context.

- Specification of a system architecture for the deployment of advanced network selection or TFAP decision schemes. Basic system requirements are: (a) availability of cross-layer (device configuration and user preferences, network and application context) information in the decision points, (b) support for dynamic participation of different network operators, (c) interoperability in the information exchange among different actors (mobile terminals, network operators, service brokers), (d) support for personalized computational intensive decision algorithms irrespective of MMTs' hardware configuration. The proposed architecture is based on software agents that are representatives of (i) users, (ii) network operators, (iii) the regulator and (iv) a service broker (Multi-Access Provider) that serves AAA and system management purposes. The thesis specifies agent functionality, interaction protocols among them and a fragment of an ontology for the representation of bearer services provided by network operators. A basic feature of the architecture is the support for personalized decision making deployed on the network side and the employment of agent mobility for minimal communication delay with the MMT.

- Evaluation of the feasibility of the agent-based approach, in terms of introduced communication overhead due to agent interaction and agent mobility, through the implementation of a simulation system. The simulation system is based on a widely used agent development framework, Java Agent Development framework (JADE), and simulation results support system feasibility.

This contribution has been originally presented in the following publications:

- Vassilis E. Zafeiris and E.A. Giakoumakis. Optimized traffic flow assignment in multi-homed, multi-radio mobile hosts. *Elsevier Computer Networks*, In Press, 2010.

- Vassilis E. Zafeiris and E. A. Giakoumakis. An agent-based perspective to handover management in 4G networks. *Wiley Wireless Communications and Mobile Computing*, 8(7):927-939, 2008.

- Vassilis E. Zafeiris and E. A. Giakoumakis. Towards flow scheduling optimization in multihomed mobile hosts. In *Proc. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC 2008)*, pages 1-5, Sept. 2008.

- Vassilis E. Zafeiris and E.A. Giakoumakis. Mobile Agents for Flow Scheduling Support in Multihomed Mobile Hosts. In *Proc. International Wireless Communications and Mobile Computing Conference, (IWCMC 2008)*, pages 261-266, 2008.

- Vassilis E. Zafeiris and Emmanuel A. Giakoumakis. An Agent-Based Architecture for Handover Initiation and Decision in 4G Networks. In *Proc. IEEE 6th International Symposium on World of Wireless Mobile and Multimedia Networks, (WOWMOM 2005)*, pages 72-77.

Chapter 2

Related Work

The related work is divided in two sections. The first one focuses on algorithms and approaches for traffic flow assignment. The second elaborates on related work on architectures for supporting the employment of flow assignment and advanced network selection algorithms.

2.1 Optimized traffic flow assignment in multi-homed hosts

2.1.1 Network selection algorithms

Network selection algorithms apply to a special case of TFAP where the MN is single-homed, i.e., it may not simultaneously activate more than one radio interfaces to serve its traffic. Thus, the problem degenerates to selection of the best available bearer service for serving application traffic on the basis of user preferences, device restrictions and traffic flows' requirements.

2.1.2 Traffic flow assignment in fixed hosts and multi-homed networks

Multi-homing is often employed by stub networks (enterprises or Internet Service Providers - ISPs) in order to enhance the reliability, performance or independence (avoiding lock-in to a single provider) of their Internet connectivity. Techniques and challenges related to the realization of multi-homing in IPv4 networks are summarized in [LX07]. Effective usage of a stub network's links with its ISPs is a complementary issue to multi-homing deployment and involves distributing incoming and outgoing traffic to appropriate links for cost and performance optimization. Such active control of traffic routing, often referred to as *smart routing*, is performed by the network's edge routers. Goldenberg et. al [GQX+04] study the assignment of traffic flows of a user, multi-homed through various ISPs, under a percentile-based charging model. This work focuses on link management and traffic assignment from the perspective of a mobile node that is multi-homed through two or more radio access networks. This perspective differentiates the aforementioned problem in terms of requirements and constraints. In contrast to a multi-homed access router, available access links for a mobile node may change over time, due to user movement to different locations. Moreover, a MMT may not be able to utilize more than one links of the same RAT (if it is equipped with a single radio interface that is compatible with that RAT), thus, an access link selection process is involved in traffic assignment optimization. As link QoS may vary across different RATs and operators, both QoS and capacity requirements of traffic flows need to be taken into account during flow assignment. This thesis incorporates power consumption into the set of objectives under optimization, due to the importance of user autonomy in a mobile networking setting.

End-host multi-homing is emerging as an effective solution for enhancing the performance and reliability of wireless access in a multi-access/multi-RAT network setting. The PERM framework is proposed towards this direction [THL06], that enables collaborative Internet access through residential WLANs. PERM resides in end-hosts, deals with non real-time traffic flows and assigns them upon their establishment to appropriate wireless links. Flow assignment is based on: (a) prediction of flows' expected traffic volume and (b) monitoring and capacity estimation of available links. PERM scheduling involves prediction of the end-host's destination IP addresses that is based on their time correlation. Thus, for each remote IP address a list of other IP addresses is maintained, sorted by the number of connections within a recent time window. Given the last visited IP address, one of the first m addresses in its associated list will be visited with high probability. Moreover, for each remote IP address the traffic volume of its last n connections is maintained as a time series and the volume of a newly arriving flow is estimated as a weighted-sum of its predecessors. Traffic flows are classified to light-volume ($<$ 8KB) and heavy-volume ones and their scheduling is based on the Round Trip Time (RTT) and throughput of available wireless links. Flow scheduling targets the minimization of either the flows' total transmission time or the maximum transmission time. Transmission time for a flow i scheduled to link j is defined as $V_i/A_{i,j}$ where V_i is the flow's expected traffic volume and $A_{i,j}$ the achievable bandwidth on the link. Moreover, $A_{i,j} = B_j/(ZRTT_i^2)$ where RTT_i is the RTT of link i and Z the total number of flows scheduled to i. Since the optimization problems are NP-complete integer non-linear problems and their parameters constantly change the author propose heuristics for their solution. This work focuses on scheduling of both real-time and non-real time flows for economic cost and power consumption optimization. The proposed problem formulation takes into account flow performance in the form of guarantees on minimum QoS requirements, rather than as objective for optimization that is the case in PERM.

2.1.3 Middleware for traffic flow assignment

Bonin et. al [BLH09] designed and implemented a middleware, named Ubique, for management of a MMT's radio interfaces and optimal assignment of application traffic flows to them. Flow assignment is based on information that is modeled and stored as a set of profiles that are relevant to user preferences, network context, application requirements etc. The middleware comprises functional entities for the management and retrieval of profile information, as well as for decisions on flow assignment. With regard to the latter, it is described as a multi-objective optimization problem without including details on the problem formulation and the objectives under optimization. This work is complementary to Ubique as it specifies an approximation algorithm for the flow assignment problem that can be incorporated as a decision method in the architecture.

Nguyen et. al [NVAGD08] adopt a two phase approach for the management of connectivity of a multi-interface host with each phase being triggered by different events. The first phase involves radio interface activation, on the basis of policies that take into account battery level, user preferences, velocity and traffic activity. Network selection is triggered, in a second phase, by: (a) changes in radio interfaces' activation status, (b) user movement towards the boundaries of currently used cell(s) and (c) changes in traffic load. Network selection is based on a utility function that consists of the weighted product of various factors such as power consumption, cost, network load and capacity, and Received Signal Strength (RSS). Access networks with highest utility are associated with the respective radio interfaces. A challenge for this approach is the appropriate selection of network selection attribute weights. Moreover, the proposed network selection procedure is more appropriate for scenarios where a single interface can serve all user traffic. In case that two or more interfaces need to be activated, a flow-to-interface assignment phase must also be incorporated. The assignment of traffic to each interface may also require the selection of different weight vectors for each interface, e.g., due to the different QoS requirements of assigned flows. In this work, the radio interface activation, network selection and traffic assignment problems are addressed in a unified manner with focus on economic efficiency and energy autonomy.

Bellavista et. al [BCG09] provide a conceptual model for connectivity management in a multi-access/multi-RAT setting. The model is based on three basic concepts: interface, connector and channel. Interfaces represent the wireless hardware equipment of a device, while connectors correspond to infrastructure or ad hoc points of access that provide connectivity. A channel refers to a pair (interface, connector) properly configured for serving user traffic. On the basis of these concepts three levels of connectivity flexibility are defined: (a) 4G, where a MMT uses exactly one of its interfaces to serve application traffic and switches among them with vertical handovers, (b) ABC, where two or more interfaces may be active at the same time, each one associated with a single connector and (c) Always Best Served (ABS) that extends ABC with the capability of an interface to use more than one connectors (multiple channels per interface).

The authors in [BCG09] propose a Mobility-Aware Connectivity (MAC) middleware for channel management in an ABS context. The middleware comprises a Context Gathering layer that incorporates components for (a) network context information retrieval (Network Interface Provider) and (b) host mobility state estimation (Mobility and Peer Estimator). This work is relevant to the Metric Application layer of the proposed middleware that evaluates and selects connectors and channels on the basis of applications' connectivity requirements. The Metric Application layer comprises the Connector Manager (CoM) and Channel Selector (ChaS) components. CoM performs evaluation of available channels, in terms of durability and reliability, on the basis of factors such as host mobility status, connector range, level of trust, energy consumption. A subset of them is provided to ChaS as candidate channels for serving application traffic flows. ChaS evaluates candidates channels with metrics expressing application requirements such as Bit Error Rate (BER), delay, jitter, bandwidth and channel durability. Selected channels are passed to CoM in order to be associated with mobile host's radio interfaces. Although, MAC focuses on selection of candidate channels it does not provide a detailed specification of the algorithm that determines the channels that will be finally utilized and the distribution of application traffic flows to them. This work is based on the conceptual model of [BCG09] and assumes an ABC level of connectivity flexibility. It contributes to this conceptual model by providing an analytical formulation of the problems handled by the CoM and ChaS components. These problems are combined into a single optimization problem with focus on cost and power consumption optimization.

2.1.4 Combinatorial optimization formulations and algorithms

Flow assignment in a multi-interface host is modeled as an optimization problem in [KAE07, XV05, GAM05]. Reference [KAE07] studies the problem from the perspective of an enterprise that seeks to assign its outbound traffic to multiple ISPs. Two types of flows constitute outbound traffic: (a) *size-fixed* flows that do not usually set constraints on transmission duration but require all their data to be transmitted and (b) *time-fixed* flows that have fixed duration with specific QoS requirements and their transmitted data size can be compressed. Size-fixed flows correspond to non real-time flows, e.g., file downloads, database transactions, and are characterized by their data volume. On the other hand, time-fixed flows represent real-time audio/video flows and are characterized by their duration, the preferred and the minimum transmission rate. Flows are served through assignment to network resources that are characterized by bandwidth, duration and quality. Quality is expressed in terms of packet loss rate that lowers the available bandwidth (effective bandwidth) of a resource. A network resource represents bandwidth capacity available for a fixed duration by an ISP and is charged with a fixed cost. The assignment of flows to resources incurs network usage cost, due to capacity reservation, and opportunity costs due to not meeting the preferred transmission rate of time-fixed flows. Thus, the flow assignment problem involves establishing a trade-off between economic and opportunity cost without violating network resource capacities and minimum transmission rates of time-fixed flows. A special case of the problem with identical resources and equal preferred and minimum transmission rates for time-fixed flows corresponds to the two-dimensional bin-packing problem (2D-BPP) that is NP-hard in the strong sense.

Reference [XV05] models flow assignment as a bin packing problem where access networks correspond to bins and flows to items that need to be packed. Each flow is characterized by bandwidth, maximum delay requirements and the capability of being partitioned across more than one radio interfaces. Moreover, each flow is associated with an access preference for assignment to a certain network. Not meeting flow access preferences incurs dissatisfaction cost. Each access network has bandwidth, maximum delay and power consumption attributes. The flow assignment problem, termed Multi-Constraint Dynamic

Access Selection (MCDAS) problem, involves minimization of power consumption and dissatisfaction costs subject to network capacity and flow QoS constraints. The problem is further categorized as online, in case that flows arrive sequentially without knowledge on the arrival of subsequent flows or offline when all flows are initially available. The problem formulation does not take into account network usage cost and assumes that both provided capacity and traffic flows correspond to a single uplink or downlink direction.

In [GAM05] the problem is formulated as a Multiple choice Multiple dimension Knapsack Problem (MMKP) with multiple knapsacks. Specifically a set of unidirectional traffic flows F_j is considered and each flow is associated with a QoS profile comprising various QoS levels, q_{jz}. Available access networks are mapped to knapsacks K_i with bounded capacity representing their bandwidth. The formulation assumes the existence of the functions: (a) $R(q)$ that maps a QoS level q to a network resource level r, (b) $U(q)$ that maps a QoS level to user utility u and (c) $C(r)$ that maps a resource level to economic cost. The assignment of q_{jz} to K_i results to user profit $u_{ij} = U(q_{ij})$ and incurs network usage cost $c_{ij} = C(R(q_{ij}))$. The traffic flow assignment problem involves packing flows, at appropriate QoS levels, to a minimal subset of knapsacks so that total net utility $(\sum u_{ij} - c_{ij})$ is maximized.

The aforementioned approaches to flow assignment ([KAE07, XV05, GAM05]) are not directly applicable to the context of a multi-homed mobile host. The main reason is that not all combinations of available wireless networks, that provide the combined capacity required by traffic flows, may be part of a problem solution. As the MMT is equipped with a finite number of radio interfaces, each one corresponding to a specific RAT and being capable of associating to a single RAN, candidate problem solutions span only the subsets of available networks that can be simultaneously bound to the MMT's radio interfaces. This constraint maps flow assignment to a generalization of the Multiple Knapsack Problem with Assignment Restrictions [DKK+00] instead of the multiple knapsack or bin packing problem.

The problem of traffic flow assignment in a multi-homed mobile network is studied in [WWY+07]. The problem context involves a network deployed on a vehicle that follows a predefined route (e.g., a bus or train). The route comprises a sequence of sites with different access network availability. The authors assume a predefined set of networks, each one having presence in one or more sites of the vehicle's journey. The mobile network serves application traffic flows by distributing them to the set of networks that are available in each site. Given that the vehicle moves between sites in predictable time intervals, the duration of each flow is defined in terms of a sub-sequence of sites. The traffic flow assignment problem involves assignment, for each site, of its active traffic flows to available access networks with the objective of minimizing handover cost. The authors study the complexity and provide heuristic algorithms (a) for online and offline versions of the problem and (b) for flows with or without capability of being partitioned across two or more radio interfaces. This traffic flow assignment problem formulation, also assumes that the mobile network may simultaneously utilize all available networks without being restricted by the number and RATs of its edge router's radio interfaces. Moreover, it does not take into account network usage costs and assumes that all available networks satisfy the traffic flows' QoS requirements.

Finally, another approach that focuses on assigning traffic flows to available connections is described in [SKK08]. The problem is called Network Connection Selection (NCS) problem and involves assigning n traffic flows to m already established connections of a MMT. Basic assumptions of the problem formulation are: (a) available connections satisfy the QoS requirements of all flows, (b) traffic flows are non real-time and have the same direction, e.g., file downloads, (c) flows are served sequentially by each radio interface, (d) the network usage cost and duration of transferring flow j through connection i are problem parameters with values c_{ij} and d_{ij} respectively. NCS is a bi-objective integer non-linear optimization problem that targets the joint minimization of the total transfer cost and the maximum duration of file transfers on all radio interfaces. The authors approximate problem solutions through a heuristic algorithm that is applied on a single-objective version of the problem. This problem version is based on a single objective function that produces a compromise solution and is a linear combination of the cost and maximum duration functions. Nevertheless, the problem formulation in [SKK08] sets enough assumptions to make it too specific for its application in an ABC context, where different types of radio interfaces are supported, network selection is required and bearer service properties as well as traffic flow requirements in terms of QoS are not identical.

2.1.5 Concurrent Multipath Transfer

A research area that is relevant to the utilization of multiple radio interfaces for increasing the performance and reliability of wireless access is Bandwidth Aggregation (BAG) that is also referred to as Concurrent Multi-path Transfer (CMT) [CR06, IAS06, LWZ08, FCCM07, KS07, TS09, KZSH05]. BAG refers to the problem of scheduling user traffic to multiple active connections in a way that packet reordering in the receiver side is minimized. The ultimate goal is to have multiple radio access connections that behave as a single one with aggregated bandwidth. Both network layer (e.g., [CR06]) and transport layer (e.g., [IAS06, LWZ08, FCCM07, KS07, TS09, KZSH05]) approaches have been proposed for BAG. Transport layer approaches focus mainly on: (a) enhancements to transport layer protocols such as TCP or Stream Control Transmission Protocol (SCTP) [SXM$^+$00] for the incorporation of CMT capability [IAS06, LWZ08, FCCM07, KZSH05] and (b) utilization of multiple radio interfaces of a single or multiple hosts for improving TCP performance in a mobile networking setting [KS07, TS09].

BAG solutions schedule traffic at the packet level and not at the flow level that is the focus of this work. Scheduling of flows trades off the fine-grained control that packet scheduling offers, for lower processing costs and ease of deployment with minimal interventions in network infrastructure or device protocol stack. Nevertheless, the problem formulation proposed in Section 3.2.1 can be extended for the support of CMT-enabled SCTP flows by relaxation of the flow integrality constraint.

2.2 Architectures enabling TFAP optimization schemes

Traffic flow assignment involves network selection for one or more radio interfaces of the MMT and decision on the distribution of application traffic flows to them. Network selection constitutes the main responsibility of the handover decision mechanism (it also focuses on determining the timing for handover execution). The support of optimized network selection in a MMT, with multi-homing capability, that operates in a heterogeneous radio access environment requires the presence of appropriate software infrastructure deployed to both the MMTs and the network side. A short review of architectures proposed for such an infrastructure is included in this subsection. The various architectures are categorized in terms of the degree of their components' distribution among the MMTs and the network side. In "host-centric" architectures, system components are deployed on end-hosts, while in "network-based" architectures their components are distributed to both MMTs and network nodes. Approaches that are specified by standardization bodies are presented separately in the first part of this section.

2.2.1 Standardization efforts

IEEE 802.21 Media Independent Handover Services

The IEEE 802.21 Media Independent Handover (MIH) Services standard focuses on enabling optimized handovers across heterogeneous access networks [IEE08]. IEEE 802.21 does not introduce a mobility management protocol but instead integrates in the mobility management protocol stack of both MNs and network entities and provides a framework for deploying advanced handover initiation and decision policies. Its services are delivered through the cooperation of MIH Function (MIHF) entities that are deployed to MNs and to access or core network elements. The IEEE 802.21 standard specifies the responsibilities of a MIHF and a protocol for the interaction of MIHFs deployed to different network entities. Moreover, it defines service access points (SAPs) for the integration of the MIHF to the protocol stack of its execution environment.

The MIHF provides services to Layer-3 and above mobility management protocols, as well as to the management and data bearer plane of a network node through a media independent handover service access point (MIH_SAP). The MIH_SAP interfaces with Layer-3 and above protocols for the support of vertical handovers, as horizontal handovers are out of scope of IEEE 802.21. The services that MIH_SAP provides access to are:

- the Media Independent Event Service (MIES) that detects changes in link layer properties and initiates appropriate events to upper layers (e.g., Layer 3),
- a Media Independent Command Service (MICS) that provides a set of commands for controlling link-layer properties relevant to handover and switching among different links if required,
- a Media Independent Information Service (MIIS) that provides information on available access networks and their service properties, thus enabling advanced handover decisions across different access technologies.

The MIHF also interacts with different types of link layers of its local network node through media dependent SAPs (MIH_LINK_SAPs) in order to deliver its services to the upper layers. Moreover, it uses the services of remote MIHFs through a media dependent transport SAP (MIH_NET_SAP). Note that the MIHF is deployed in MNs (MN-based MIHF), to Point of Access (PoA) network elements (e.g., 3GPP base stations, WLAN Access Points) and to non-PoA core network elements of the various networks. A MIH

Network Entity that interacts with a MN-based MIHF acts as its MIH Point of Service (MIH PoS). However, it is possible that a MIH Network Entity may not interact with MN-based MIHFs but instead provide services to MIH Pos elements and in this case it is called MIH Non-PoS element.

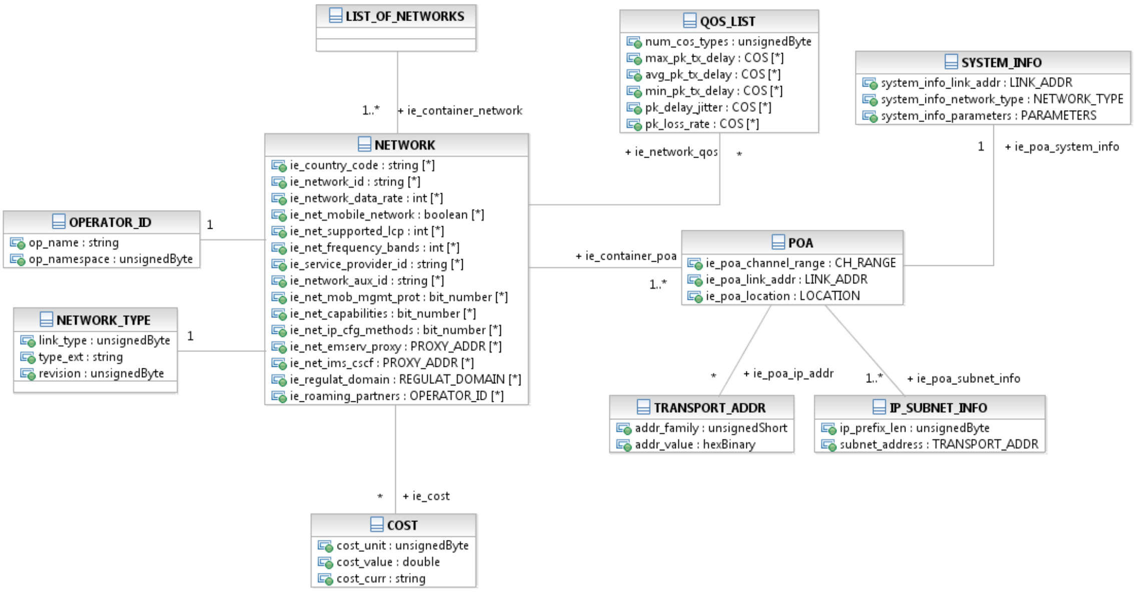
LIST_OF_NETWORKS
1..*
+ ie_container_network
QOS_LIST
num_cos_types : unsignedByte
max_pk_tx_delay : COS [*]
avg_pk_tx_delay : COS [*]
min_pk_tx_delay : COS [*]
pk_delay_jitter : COS [*]
pk_loss_rate : COS [*]
SYSTEM_INFO
system_info_link_addr : LINK_ADDR
system_info_network_type : NETWORK_TYPE
system_info_parameters : PARAMETERS
+ ie_network_qos
*
1
+ ie_poa_system_info
NETWORK
ie_country_code : string [*]
ie_network_id : string [*]
ie_network_data_rate : int [*]
ie_net_mobile_network : boolean [*]
ie_net_supported_lcp : int [*]
ie_net_frequency_bands : int [*]
ie_service_provider_id : string [*]
ie_network_aux_id : string [*]
ie_net_mob_mgmt_prot : bit_number [*]
ie_net_capabilities : bit_number [*]
ie_net_ip_cfg_methods : bit_number [*]
ie_net_emserv_proxy : PROXY_ADDR [*]
ie_net_ims_cscf : PROXY_ADDR [*]
ie_regulat_domain : REGULAT_DOMAIN [*]
ie_roaming_partners : OPERATOR_ID [*]
OPERATOR_ID
op_name : string
op_namespace : unsignedByte
1
NETWORK_TYPE
link_type : unsignedByte
type_ext : string
revision : unsignedByte
1
+ ie_container_poa
1..*
POA
ie_poa_channel_range : CH_RANGE
ie_poa_link_addr : LINK_ADDR
ie_poa_location : LOCATION
*
+ ie_poa_ip_addr
1..*
+ ie_poa_subnet_info
TRANSPORT_ADDR
addr_family : unsignedShort
addr_value : hexBinary
IP_SUBNET_INFO
ip_prefix_len : unsignedByte
subnet_address : TRANSPORT_ADDR
*
+ ie_cost
COST
cost_unit : unsignedByte
cost_value : double
cost_curr : string

Figure 2.1: MIIS Basic Information Schema

The MIH enables the definition and deployment of advanced handover initiation and decision policies with varying levels of control distribution among the MNs and network nodes. Thus, it supports the definition of Mobile Controlled Handover (MCHO) policies, Network Controlled Handover (NCHO) and Mobile Assisted Handover (MAHO) policies. Handovers may be initiated by local or remote events received through subscription to the local MIHF or to a remote MIHF of a MIH PoS. Handover decision is based on static and dynamic information related to the available networks in the MN's current location. Static information spans link-layer parameters such as channel information, MAC address and security information of the PoA, QoS and cost information, as well as information on higher layer services provided through the PoA. Such information is provided by MIIS servers through appropriate queries that may include RAT, operator and geographical range criteria. Figure 2.1 presents the main information elements of the MIIS schema that includes network, network type, PoA and operator related concepts. Dynamic information on resource availability of PoAs can be directly retrieved through the MICS of their serving MIH PoS. Thus, MIH allows a MN to receive information on available access networks through a single radio interface. However, after the handover decision the MN has to actively scan and measure through its radio interfaces the signal strength of candidate access networks in order to verify their capability of providing the required QoS and capacity.

3GPP EPS Access Network Discovery and Selection Function

3GPP Release 8 EPS has introduced a core network element for supporting a MN in efficient access network discovery and selection in a multi-RAT multi-access radio access environment. This EPC element is called Access Network Discovery and Selection Function (ANDSF) and it delivers its services to network subscribers with the role of Home-ANDSF (H-ANDSF) or to roaming users with the role of Visited-ANDSF (V-ANDSF) [3GP09a]. Although ANDSF services are not required for a MN's base operation, they contribute to better adaptation to network context and user preferences.

ANDSF services are employed by the MN's handover decision function and contribute to handover decisions aligned with operator policies and user preferences. ANDSF's role involves: (a) provision to the MN of the operator's inter-system mobility policy, (b) delivery, upon MN request, of the list of available access networks on the basis of RAT and location-based criteria. The inter-system mobility policy is updated either by network triggers or after MN request for network discovery and selection information. ANDSF supports mobile-assisted vertical handovers as the decision on the target network for handover execution is derived by the cooperation of the home/visited EPS network and the MN. Specifically, ANDSF transfers to the MN operator policies and network availability information that determine the candidate access networks for handover, while the MN selects appropriate candidate networks on the basis of user preferences.

ANDSF is deployed in the core network of a 3GPP EPS network and delivers its services through direct communication with the MN. MN interaction with ANDSF takes place over a secure IP connection and is based on the S14 reference point [3GP10b]. Both push and pull methods are used for information transfer over S14 that is based on the Open Mobile Alliance Device Management (OMA DM) protocol and the respective management object (MO) management object defined in 3GPP TS 24.312 [3GP10a]. The ANDSF MO defines the structure of the policies and the discovery information. Each policy comprises a set of rules that define the priority of access networks. Moreover, each rule is associated with geographical and temporal information that determine the conditions that the rule is valid. The discovery information includes the type and the service area (geographical area) of an access network as well as access specific information such as channel information, cell type etc. More elaborate information is left to be defined by vendors and standardization bodies of the respective access technologies (e.g., OMA-DDS_ConnMO-V1_0 [OMA08a] and OMA-DDS_ConnMO-WLAN [OMA08b] for WLAN related information).

IEEE 1900.4 standard

The IEEE 1900.4 standard [IEE09a] defines a management system, its architectural elements and the information exchanged among them, for distributed optimization of radio resource usage in a heterogeneous radio access environment. This environment comprises multiple MNs and Composite Wireless Networks (CWNs). A CWN is defined in [IEE09a] as a network composed of multiple radio access networks that are interconnected through a packet-based core network with IEEE 1900.4 entities deployed in it. With regard to MNs, they incorporate multiple radio interfaces with or without multi-homing support. IEEE 1900.4 focuses on the building blocks, distributed in mobile terminals and CWNs, for context-awareness, generation and enforcement of reconfiguration policies for optimized radio resource management. Moreover, it identifies three use cases for the management system and describes their realization through the collaboration of the aforementioned building blocks.

The system requirements for IEEE 1900.4 are represented by the use cases: (a) dynamic spectrum assignment, (b) dynamic spectrum sharing and (c) distributed radio resource optimization. The first use case refers to the dynamic assignment of frequency bands to RANs within a CWN, operating in a given geographical area and time, for spectrum usage optimization. Dynamic spectrum sharing involves different RANs and terminals gaining dynamic access to fully or partially overlapping spectrum bands in a way that causes less than an admissible level of mutual interference. Finally, distributed radio resource optimization refers to allocation of radio resources for serving traffic needs in a way that network, terminal device and user objectives are satisfied. The latter assumes static assignment of channels to RANs and is more relevant to this thesis.

In IEEE 1900.4, the decision making involved in radio resource optimization is distributed among CWNs and MNs. Specifically, the network side is responsible for generation and transfer of appropriate policies, while the MNs decide on the assignment of their traffic flows to available radio resources in a way that network policies are not violated. Enforcement of MN decisions may involve its reconfiguration or execution of handovers, procedures that are out of scope of this standard as concerning their realization. Note that terminal reconfiguration refers to reconfiguration of its hardware, and/or software in order to change its operating parameters in the physical or link layers (e.g., carrier frequency, signal bandwidth, radio interface) in one or more of its radio interfaces [IEE09a]. The decision points in the IEEE 1900.4 architecture are the Network Reconfiguration Manager (NRM) and the Terminal Reconfiguration Manager (TRM) deployed in the core network of CWNs and MNs respectively. Possible deployment options for the NRM (and the IEEE 1900.4 architecture as well) are: (a) single CWN with a single NRM node, (b) multiple CWNs with a common NRM and (c) multiple CWNs, each one owning its NRM that collaborates with NRMs of others. With regard to TRM, each MN has its own instance of TRM.

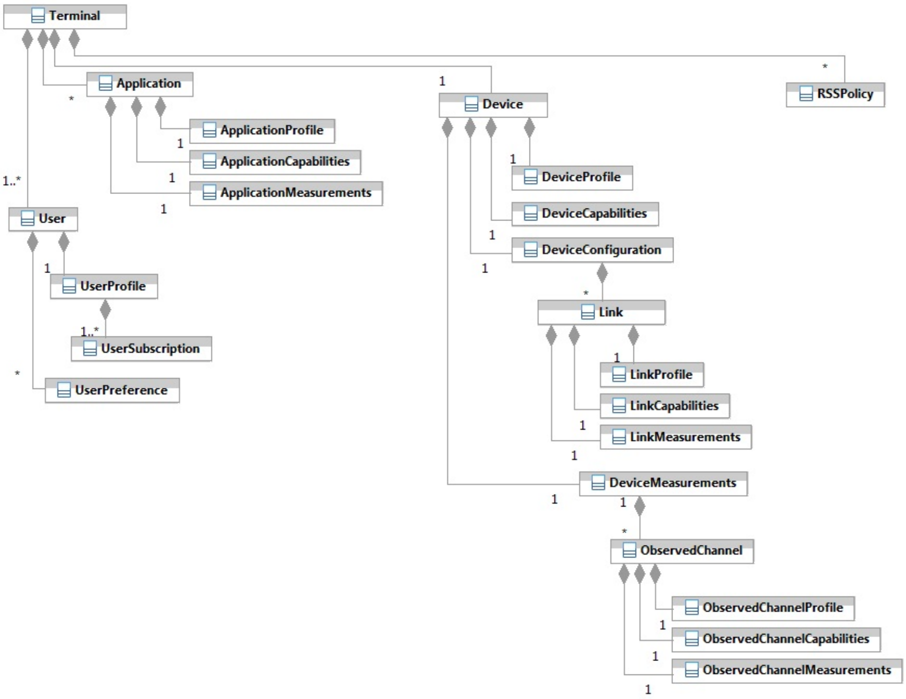

Figure 2.2: Terminal related classes.

NRM collects network context information from all RANs connected to its CWN, as well as from RANs of other CWNs through the respective NRMs. Moreover, it collects terminal context information from the TRMs of MNs connected to its CWN or from other NRMs for terminals served by other CWNs. Its decision making is also based on spectrum assignment policies that are provided by the Operator Spectrum Manager (OSM) building block deployed in each CWN. NRM generates radio resource selection policies that are propagated to the TRMs of mobile terminals. On the basis of these policies, network and local context information, each TRM decides on the radio resources that will be used for serving its traffic. The enforcement of TRM decisions is handled by the Terminal Reconfiguration Controller (TRC) block that is also deployed in each MN. NRM may also generate RAN reconfiguration requests that are handled by the RAN Reconfiguration Controller (RRC) element available in each RAN.

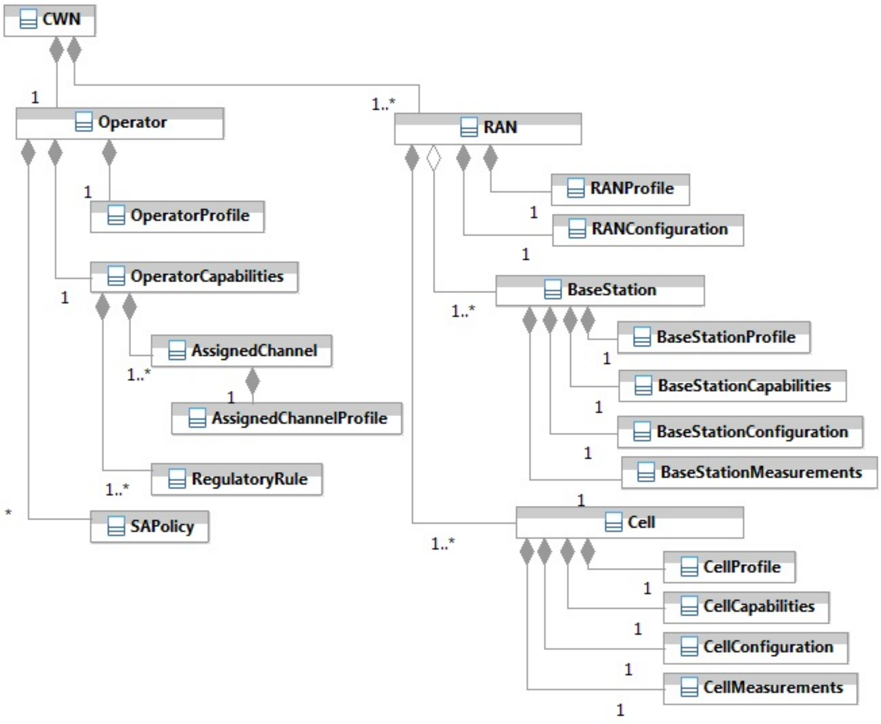

Figure 2.3: CWN related classes.

Policy generation and decision making is based on rich network and terminal context information that is retrieved by appropriate IEEE 1900.4 blocks deployed in RANs and MNs, namely, RAN Measurement Collector (RMC) and Terminal Measurement Collector (TMC) respectively. Figures 2.2 and 2.3 present UML class diagrams with the main information elements comprising network and terminal context.

IEEE 802.11u amendment for Interworking with External Networks

The IEEE 802.11u draft standard specifies an advertisement service that enable faster discovery and selection of WLANs [IEE09b]. Generic Advertisement Services (GASs) provide transport mechanisms for advertisement services toward mobile hosts that are either associated or not with an Access Point (AP). IEEE 802.11u enables a disconnected MN to discover information related to services provided by an infrastructure WLAN, as well as services accessible through external networks that interwork with the WLAN. GAS is designed to support multiple query protocols and allows a mobile terminal to access information available locally to the AP or to retrieve it from external networks, e.g., it enables access to IEEE 802.21 MIIS. Description of an information model for advertisement services is out of scope of IEEE 802.11u.

2.2.2 Host based solutions

A layered middleware architecture is proposed in [BCCF05] for supporting the development of mobility-aware and context-aware applications. The middleware comprises two layers: the Mechanisms layer and the Facilities layer. The Mechanisms layer includes RAT-dependent modules and specifically one module for each supported RAT by the MN. Each module includes a context gathering component and a component for managing the handover procedure. Vertical handovers are not handled at the Mechanisms layer but at the Facilities layer that coordinates the operation of the RAT specific modules. The latter comprises two general purpose facilities, namely NCSOCKS and Mobility Awareness & Management (MM), and one domain-specific Multimedia Streaming facility. The MM facility includes a Connection Monitor component that gathers context information on available connections and a Location monitor component that retrieves the MN's location. Their focus is on providing a uniform context representation to the services layer. Moreover, the MM facility manages the vertical handover procedure. The facilities layer provides an API for applications to control terminal connectivity on the basis of context and their connectivity requirements. However, this approach results in an application specific management of the handover procedure, while a mechanism for resolving possible conflicts, in case that multiple applications are simultaneously contesting for connectivity, is not provided .

In [SRJS05] an end-to-end middleware is proposed that provides applications with transparency from changes in connectivity through a channel abstraction. Each channel is associated with a transport layer connection and is constantly monitored and managed by a Channel Management Agent (CMA) executing in each end-host. An advantage of such architectures is their support of handover management without modifying the existing networking infrastructure. However, legacy applications need to be rewritten in order to utilize their features. Moreover, the overhead related to handover management is distributed to the usually resource constrained MNs.

Tramcar (Transport and Application Layer Architecture for vertical Mobility with Context-awareness) is a host-based solution for vertical handover control [HNH07]. It targets mobile hosts with multiple radio interfaces that employ a transport layer solution for mobility management. Specifically, Tramcar controls the execution of an SCTP variant for mobility management, mobile SCTP (mSCTP) [RT07], by modifying the set of IP addresses that are available for active connections and choosing the primary address after handover decision. The proposed architecture comprises a Handover Manager (HM) and a Connection Manager (RM) component. HM evaluates available access networks and selects, when required, the most appropriate for handover execution. Access network selection is based on ranking each network through an objective function that combines in a weighted sum the normalized values of its cost, service (available bandwidth, reliability) and power consumption attributes. The CM is responsible for (a) network discovery, (b) registration to networks that are handover candidates and (c) handover execution. Networks that are handover candidates are determined by HM and CM configures an IP address to each one of them. These addresses are added to active SCTP connections as secondary addresses. Once a decision for handover execution is made by HM, CM informs the correspondent node of the new primary address. Although Tramcar's deployment does not require changes in the network infrastructure its applicability is limited to cases where both communicating endpoints support an SCTP variant for mobility management. Moreover, the actual mechanism for context information retrieval and the structure of this information are not specified.

2.2.3 Network-supported architectures

This category of approaches for support of optimized handovers in a heterogeneous radio access environment is characterized by distribution of the decision mechanism and the architecture components between user terminals and the network.

In [AMX05] an Architecture for Ubiquitous Mobile Communications (AMC) is proposed for the integration of heterogeneous communications' systems. The concept of Network Inter-operating Agent (NIA) is introduced that handles AAA and inter-domain mobility management across various wireless providers. NIA is managed by a third-party that maintains Service-Level Agreements (SLAs) with the network operators. Wireless networks integrate with the NIA through Interworking Gateways (IG). AMC supports a two level handover decision process that involves both the NIA and the MNs. Initially, the MN performs network selection on the basis of factors such as network conditions, user and application requirements, while in a second level this decision is adjusted by the NIA that focuses on global load balancing.

PROTON [VBS$^+$05] is an autonomic system for context-aware mobility management in a heterogeneous 4G network setting. Handover decision in PROTON is based on Event-Condition-Action policies that are transformed into deterministic finite state automata before their transfer to the MN for evaluation and enforcement. PROTON components are deployed in end-hosts and the network side. Network side components comprise tools for policy creation and a repository for storing policies and their respective finite state automata (FSA). The latter are transferred to MNs through a Model Deployment module. An end-host includes a Context Management Layer (CML) that perceives context events and makes them available to a Policy Management Layer (PML). PML is responsible for control and evaluation of policies that determine the MN's behaviour. Specifically, based on feedback from CML, PML retrieves from a local repository an appropriate FSA (that represents a set of policies), and produces, through its evaluation, possible actions for execution. Actions are processed by an Enforcement Layer that is responsible for execution of handovers. PROTON allows the evaluation of complex policies even in mobile devices with limited processing power capabilities. Its only requirement is enough memory for storing FSAs that represent policies. However, a limitation of this architecture lies on the fact that collected context is restricted by the MN's context perceiving capabilities, while third-party information is not utilized. Moreover, it is not clear how this approach scales in terms of policy complexity in a multi-homed mobile terminal that needs to serve multiple traffic flows of different types. In this case possibly different automata may be required for each available radio interface and a conflict resolution mechanism may be needed.

In [WFP$^+$06] a context-aware framework for handovers is proposed that introduces repositories for collection of context information and an execution platform for the dynamic deployment and execution of context handling components. The framework is based on active-networking, i.e., its components are deployed in a programmable platform installed on network nodes and mobile nodes. The proposed architecture comprises the context-gathering components, as well as a service deployment framework for the dynamic deployment and update of these components when required. Context information is stored in a series of network repositories, namely, Location Information Server (LIS), Network Traffic Monitor (NTM) and User Profiles' Repository. LIS tracks and provides location information related to MNs and APs while NTM monitors the available bandwidth of APs. The context management framework includes two types of entities: (a) Context Collection Points (CCPs) that are deployed in the network side and serve the collection and filtering of context information available in the various repositories, (b) Handover Decision Points (HDPs) that consume context information from CCPs for network selection purposes. In the proposed handover management architecture MNs have the role of HDPs, while CCPs also have the role of the Handover Manager that controls the handover procedure from the network side. Both types of entities host a platform for dynamic deployment and execution of handover management modules that are retrieved from a Service Deployment Server. These modules constitute the Handover Decision Module (HDM), that implements a network selection scheme, and the Handover Support Module (HSM) that implements a context exchange protocol. HDM is deployed in MNs, while HSM is deployed to both MNs and CCPs. On the basis of this architecture the handover process involves preparation of HDPs and CCPs, if required, with appropriate modules and then network selection performed by the MN.

In [CGG08] an agent-based system is proposed for network selection in next generation heterogeneous networks. The system is called Living Systems Autonomic Service Access Management suite (LS/ASAM) and focuses on optimization of both client connectivity and radio resources of network operators. LS/ASAM's components are software agents deployed to MNs and access nodes or network management facilities of RANs. The client component is the Connection Agent (LS/CA) that is responsible for continuity and adaptation of connectivity of the MN it is deployed to. LS/CA triggers vertical handovers and performs network selection on the basis of offers provided by network-side components of LS/ASAM.

LS/Service Access Manager (LS/SAM) is a software agent deployed in an operator's network and serves load-balancing and congestion recovery purposes. LS/SAM pro-actively monitors traffic load and available resources in the access node(s) it controls and triggers handovers or session drops. Moreover, it responds to requests from LS/CAs regarding bearer service offers by taking into account operator policies, available network resources and balance of load in other available access segments. Collaboration of LS/CAs and LS/SAMs is based on the contract-net protocol [FIP02b]. System description, as presented in [CGG08], does not elaborate on the structure of a LS/CA request for resources (that may impact system responsiveness due the volume of uploaded information) as well as the way that LS/SAMs coordinate in order to load balance the RANs attached to an operator's core network.

A comprehensive framework for the exchange and management of context is proposed in [Amb05]. In this approach, shared ontologies play a key role in the representation and exchange of context information. This feature is also inherent in an agent-based approach enhancing thus information interoperability. Moreover, well-established Agent Communication Languages (ACLs) [LFP99] provide a messaging framework for context exchange among system components. A more elaborate discussion on these merits of software agents is included in Section 4.3.1.

Chapter 3

Optimized Traffic Flow Assignment in Multi-Homed, Multi-Radio Mobile Hosts

3.1 Introduction

Wireless Internet access is continuously expanding its geographical reach through a variety of Radio Access Technologies (RATs). However, no single RAT may completely satisfy the bandwidth and QoS requirements set by current and emerging multimedia applications. Moreover, each RAT focuses on different degrees of user mobility in terms of speed and range. In order to benefit from radio access diversity, many modern mobile communication devices are multi-mode, i.e., they are equipped with multiple radio interfaces (3GPP, 802.11a/b/g/n etc). Moreover, special purpose mobile devices are emerging that provide aggregated bandwidth capacity to mobile business or vehicular users, through multiple wireless broadband subscriptions [Mus09].

A Mobile Multi-mode Terminal (MMT) that is multi-homed has at least two global IP addresses, associated with respective radio interfaces. The IETF Mobile Nodes and Multiple Interfaces in IPv6 Working Group (MONAMI6 WG) has identified the benefits that mobile host multi-homing offers to both end users and network operators [EMWK08] and its successor, IETF Mobility EXtensions for IPv6 Working Group (MEXT WG), is working towards enhancing MIPv6 with multi-homing support. This capability will be enabled by allowing the registration of multiple CoAs with a certain HoA [WDT$^+$09]. Moreover, this specification is being complemented with support for binding flows to specific CoAs [STM$^+$10], allowing, thus, the execution of handovers at a traffic flow level.

Given these enhancements to a basic mobility management protocol such as MIPv6, a MMT extends its degrees of freedom for adapting its connectivity status to the changing traffic requirements and wireless networking context. For instance, the range of options for responding to the arrival of a traffic flow, when spare capacity in active radio interfaces is not available, may include: (a) activation of an inactive radio interface and its attachment to an appropriate radio bearer service, (b) horizontal handover on an active radio interface towards a higher capacity bearer service, (c) redirection of one or more traffic flows, already served by one interface, to another interface for best utilization of available bandwidth capacity and so on. The set of available options on each occasion depends on the wireless context and the MMT's traffic load and hardware configuration. Moreover, each alternative may have different impact on the fulfillment of user preferences and especially on economic efficiency and energy autonomy. Thus, evaluation and determination of the optimal operational state requires advanced and fast executing decision algorithms. Execution efficiency is required due to the frequently occurring triggers for decision making that include changes in served traffic, network conditions and device status (e.g., battery lifetime).

This section focuses on the problem of joint management of traffic, wireless connectivity and power consumption in the context of a multi-homed MMT operating in a dynamic environment. The MMT may have the role of an end-host serving its own traffic or the role of a mobile router that acts as an Internet gateway to a personal area or vehicular network. The section focuses on the problem of assignment of application traffic flows (either inbound or outbound) to appropriate radio interfaces and radio bearer services in a way that: (a) satisfies the traffic flows' QoS requirements and the bearer services' capacity constraints, and (b) establishes the best trade-off between economic cost and power consumption. For brevity reasons, the problem will be referred to as Traffic Flow Assignment Problem (TFAP). The economic cost factor of TFAP corresponds to network usage cost, while power consumption is due to the operation of active radio interfaces. Due to the dynamic nature of problem parameters the MMT iteratively faces TFAP instances of variable size during its operation lifetime.

An analytical formulation for the TFAP is provided in this section that maps the problem to a bi-objective combinatorial optimization problem. The formulation takes into account additional constraints, as compared to prior work, that contribute to a more realistic representation of the problem domain. The bi-objective problem is solved by setting one of the objectives as a target for minimization (economic cost) and the other objective (power consumption) as an additional problem constraint by appropriately choosing an upper limit for its allowed values. A study of TFAP's complexity is also provided that proves its hardness through transformation from the Multiple Knapsack problem with Assignment Restrictions (MKAR) that is NP-Hard [DKK+00]. Given the complexity of TFAP and requirements for frequent and fast execution, a heuristic approximation algorithm is introduced that is based on local search and establishes a good balance between solution quality and execution time. A basic feature of the proposed algorithm is the combined use of two objective functions that guide the search towards minimum cost solutions with an upper limit on power consumption. Solution quality is evaluated by comparing heuristic and exact solutions for a large number of randomly generated problem instances. Moreover, the approach is evaluated with a discrete event simulator in order to study its merits over the time domain and the incurred mobility management overhead.

3.2 Traffic Flow Assignment Problem - TFAP

3.2.1 Problem formulation

Assume a MMT that is equipped with a set of different technology radio interfaces, e.g., 3GPP, IEEE 802.11x, WiMAX etc. Let $\mathbf{R} = \{r_i : 1 \leq i \leq m, m \in \mathbb{N}^*\}$, be the MMT's available radio interfaces. Moreover, the MMT has multi-homing support, i.e., it is capable of simultaneously using two or more of its radio interfaces for serving its data traffic.

The MMT's current location lies on the overlapping service areas of a set of Radio Access Networks (RANs). The RANs correspond to different RATs and are managed by one or more network operators. Each RAN provides wireless access through a set of bearer services, i.e., data transfer services characterized by different Quality-of-Service (QoS) guarantees. Let $\mathbf{B} = \{b_j : 1 \leq j \leq n, n \in \mathbb{N}^*\}$ be the set of bearer services provided by the RANs that serve the MMT's current location. Each bearer service b_j is characterized by a set of service, cost and power consumption attributes. The service and cost attributes of b_j are described below, while its power consumption attributes will be described thereafter:

· $k_{u,j}$ is uplink bandwidth capacity in kbps offered by the service to the MMT,

- · $k_{d,j}$ is downlink bandwidth capacity in kbps,
- · Q_j is a QoS class that characterizes the service's performance and defines upper limits for delay, jitter and bit error rate. In this work the QoS classes defined by 3GPP [3GP08b] are adopted and, thus, the domain of Q_j is Background (Q_B), Interactive (Q_I), Streaming (Q_S), and Conversational (Q_C). Moreover, a strict total order relation among them is assumed, i.e., $Q_B < Q_I < Q_S < Q_C$ denoting that QoS requirements set by Q_B are lower than those of Q_I and so forth.
- · c_j is cost per kbit of transferred data either in the uplink or downlink direction.

A volume-based charging model is adopted due to its simplicity and fairness for both users and network operators that act in a competitive wireless access environment where network selection is enabled at the time granularity of a service session. A similar charging model is utilized today for mobile access in roaming scenarios, where a user, at any time instance between successive service sessions, is capable of selecting the services of any operator that has roaming agreements with its home operator.

The wireless access resources represented by each bearer service b_j refer to the RAN's capacity availability per admitted user in the MMT's current service area. Bearer service descriptions are retrieved by the MMT in a RAT specific manner from its current point of access or through a media independent mechanism as the Media Independent Information Service (MIIS) specified in [IEE07]. Service availability may change dynamically due to variations of cell load or radio propagation conditions. The MMT perceives these variations either directly (for currently used services) or indirectly by the RANs' information services.

Each radio interface r_i supports a set of RATs and may be associated with at most one bearer service of compatible RAT. Once associated with a bearer service, a radio interface can serve user traffic by utilizing the service's available capacity, and contributes to the MMT's overall power consumption. Its power consumption depends on factors such as: (a) RAT of the bearer service, e.g., a 3GPP radio interface has different power consumption when associated with a 2.5G or a 3G bearer service, (b) configuration of the bearer service's network point of access, e.g., support or not of power-saving mode in 802.11 APs, (c) volume and direction of served traffic. The power consumption model adopted in this thesis has been proposed in [LN03] and used in a similar context in [XV05]. Since the power consumption coefficients of this model depend on the RAT, and a radio interface may support multiple RATs, they are used in the TFAP problem formulation as bearer service attributes. Thus,

Table 3.1: Examples of power consumption parameters.

RAT	Parameter type		
	Base (μW)	Transmission (μW/kbps)	Reception (μW/kbps)
UMTS	107.2	8.3	4.9
GPRS	313.47	0.76	0.36
IEEE 802.11a	368	0.32	0.14
IEEE 802.11b	262.7	1.22	1.22

the power consumption attributes of a bearer service b_j are:

- wr_j, power consumption per kbps of received data (W/kbps),
- wt_j, power consumption per kbps of transmitted data (W/kbps),
- wb_j, power consumption (W) due to base operation of the radio interface without transferring data.

The power consumption in Watts of a radio interface that sends and receives data with rates bw_t and bw_r kbps respectively through bearer service b_j is $P = wb_j + bw_t\ wt_j + bw_r\ wr_j$. Table 3.1 includes example values of power consumption parameters for four RATs as estimated and used in [XV05], [LN03].

The set $\mathbf{B}$ of available bearer services can be partitioned into $m = |\mathbf{R}|$ disjoint subsets, i.e., $\mathbf{B} = B_1 \cup B_2 \ldots \cup B_m$, where each subset B_i comprises services that are compatible with radio interface r_i.

Definition 1. *A radio interface activation B' represents the association of one or more radio interfaces of the MMT with compatible bearer services in order to serve user traffic. B' is a subset of* $\mathbf{B}$ *with* at most *one element from $B_1, B_2, \ldots B_m$, i.e., $B' \subseteq \mathbf{B}, |B'| \leq m$. The inequality corresponds to the case where one or more radio interfaces of the MMT are deactivated.*

The presence of an element from B_i in a radio interface activation B' denotes that radio interface r_i is activated[1]and associated with that bearer service. On the other hand, the absence of an element from B_i denotes that r_i is deactivated. Given the set of available bearer services $\mathbf{B} = B_1 \cup B_2 \ldots \cup B_m$, the number of possible radio interface activations is $(|B_1|+1)(|B_2|+1)...(|B_m|+1)-1$. This accrues from the $|B_i|+1$ possibilities for each radio interface (i.e., $|B_i|$ alternative bearer services plus the possibility of being deactivated) with the exclusion of the case of all interfaces being deactivated.

The data traffic served by the MMT is modeled in terms of uplink and downlink traffic flows that are generated by user applications. Let $\mathbf{F} = \{f_z : 1 \leq z \leq v, v \in \mathbb{N}^*\}$ be the set of traffic flows that are served by the MMT at a given time instance. A flow f_z, that may be uplink or downlink, is characterized by a set of attributes, i.e., $f_z = (bw_{u,z}, bw_{d,z}, q_z)$. Specifically, $bw_{u,z}$ represents the flow's required bandwidth capacity in kbps in the uplink, while $bw_{d,z}$ states the respective requirements in the downlink. For an uplink flow f_z it holds that $bw_{d,z} = 0$, while for a downlink one $bw_{u,z} = 0$. Regarding q_z, its value represents a QoS class and has the same domain with the respective attribute of a bearer service.

Each flow f_z must be assigned to exactly one activated radio interface in order to be served. The current problem formulation does not assume flows that use a Concurrent Multi-path Transfer (CMT) protocol (like SCTP) for their transport and, thus, their traffic may not be split across two or more activated radio interfaces. A possible extension for the incorporation of this feature is the association of each flow f_z with a set of coefficients a_{iz}, where i ranges over the set of available radio interfaces. Each coefficient a_{iz} represents the fraction of flow traffic that is assigned to a radio interface r_i. Given an interface activation B', it holds that $a_{iz} \in (0, 1]$ for activated radio interfaces and $a_{iz} = 0$ for deactivated ones. The assignment of a CMT-based flow f_z, given a radio interface activation B', involves fixing the values of a_{iz} in a way that $\sum_i a_{iz} = 1$. The bandwidth capacity occupied by f_z on r_i is $a_{iz}b_z$, where b_z represents the flow's bandwidth requirements. Regarding non CMT-based flows, their respective coefficients are allowed to take only integer values, thus $a_{iz} \in \{0, 1\}$. A more detailed study of this feature and its implications on problem complexity will be part of future extensions of this work.

Definition 2. *A traffic flow assignment S with respect to the sets $\mathbf{R}$, $\mathbf{B}$ and $\mathbf{F}$, comprises a radio interface activation $B' = \{b_1, b_2, \ldots b_k\}, k \leq m$ and a partition of F into k foreign subsets $\mathbf{F} = F_1 \cup F_2 \cup \ldots \cup F_k$, such that the flows of each F_j are served by a corresponding $b_j \in B'$ without violating the bearer's capacity constraints and the flows' bandwidth and QoS requirements.*

Given a radio interface activation B', the partition of $\mathbf{F}$ into subsets must be appropriately selected in a way that for each F_j served by a $b_j \in B'$:

---

[1] The term *activated radio interface* will be henceforth used to refer to both a bearer service $b_j \in B'$ and its corresponding radio interface.

· uplink capacity of b_j is not exceeded

$$\sum_{f_z \in F_j} bw_{u,z} \leq k_{u,j}, \tag{3.1}$$

· downlink capacity of b_j is not exceeded

$$\sum_{f_z \in F_j} bw_{d,z} \leq k_{d,j}, \tag{3.2}$$

· QoS requirements of flows are satisfied

$$q_z \leq Q_j, \forall f_z \in F_j. \tag{3.3}$$

The economic cost of a traffic flow assignment S is equal to the sum of the costs incurred due to the operation of each one of the MMT's activated radio interfaces. Let $C_M(S)$ be the function that returns the economic cost of a flow assignment S and B' the radio interface activation that corresponds to S. Furthermore, let F_j be the set of flows served by each $b_j \in B'$. The economic cost of each activated interface is equal to the cost c_j of its associated bearer service b_j multiplied by the total bandwidth capacity reserved by its served flows F_j. Note that $C_M(S)$ is measured in monetary units per second and represents the incurred economic cost for each second of MMT operation under flow assignment S. The function $C_M(S)$ is defined as:

$$C_M(S) = \sum_{b_j \in B'} c_j \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}). \tag{3.4}$$

The power consumption of a traffic flow assignment S is defined in a similar manner and is given by equation 3.5:

$$C_P(S) = \sum_{b_j \in B'} (wb_j + wr_j \sum_{f_z \in F_j} bw_{d,z} + wt_j \sum_{f_z \in F_j} bw_{u,z}). \tag{3.5}$$

According to equation 3.5, the power consumption of each activated radio interface $b_j \in B'$ is equal to the sum of three addends: (a) power consumption due to base operation of the radio interface, (b) power consumption due to data reception of its inbound flows, and (c) power consumption due to data transmission of its outbound flows respectively.

Definition 3. *Given three sets* $\mathbf{R}$, $\mathbf{B}$ *and* $\mathbf{F}$ *that correspond to available radio interfaces of a MMT, available bearer services and traffic flows that need to be served, the Traffic Flow Assignment Problem (TFAP) involves finding a flow assignment* S *that minimizes both economic cost* $C_M(S)$ *and power consumption* $C_P(S)$.

The TFAP is, therefore, an optimization problem with two objectives: economic cost C_M and power consumption C_P. These objectives are independent and usually conflicting, especially in cases that the MMT serves high traffic load. Assume that the MMT's radio interfaces have access to bearer services of comparable cost and there exists a positive correlation between cost and provided capacity. If the MMT's traffic load cannot be served by a single interface (due to capacity or QoS restrictions) then optimization of C_M involves distributing traffic flows to radio interfaces with access to the cheapest bearer services. On the other hand, optimization of C_P requires the activation of the least possible number of radio interfaces by associating them with high capacity and usually higher cost bearers. Thus, TFAP is a Multi-Objective Optimization Problem (MOOP) and, depending on the problem instance, may have more than one Pareto optimal solutions (traffic flow assignments) S_p.

3.2.2 TFAP solution method

Various methods have been proposed for solving MOOPs [Coe00]. A widely used method that transforms the MOOP to a single objective problem is the weighted-sum method. The method involves prioritization of problem objectives by associating them with appropriate weight values. Then, a single objective function is defined by the weighted sum of the individual objective functions. Despite the method's simplicity, the selection of objective weights may prove a challenging issue especially in cases where: (a) objective values are not expressed in the same measurement unit and (b) measurements units do not directly reflect user utility. In TFAP, economic cost C_M is the primary user objective and is expressed in monetary units, a common measure of user utility. On the other hand, C_P is expressed in Watts and is related to energy autonomy of the MMT. Assessing the economic value of energy autonomy of specific duration is a difficult task.

These characteristics of TFAP objectives lead to the application of the ε-constraint method for solving the TFAP [Coe00]. This method also transforms a MOOP to a single-objective problem and its application involves selection of one objective (the most preferred) for optimization and introduction of one problem constraint for each one of the remaining objectives. The resulting single-objective problem is then solved for various bounds on problem constraints in order to generate the set of Pareto optimal solutions. In the case of the TFAP, the primary objective is the economic cost C_M, while power consumption can act as problem constraint. Although MOOP solving methods focus on the generation of a set of Pareto optimal solutions, the TFAP solution procedure targets a single solution, that is the economically most efficient flow assignment that does not violate a power consumption limit $P_{\max}$. This is due to the fact that users are mainly concerned on economic cost, as soon as their autonomy, till next battery recharge, is ensured.

Concerning the range of values for $P_{\max}$, it depends each time on the user context which can be classified into three basic categories:

1. "Infinite" energy resources ($P_{\max} \to \infty$), where the MMT is either plugged into an energy source or an energy source is easily accessible, e.g., when the user is at home. In this case the power consumption cost can be ignored by the user and the TFAP becomes a single objective, economic cost minimization problem, that will be referred to as TFAP$_M$.

2. Depleting energy reserves ($P_{\max} \to 0$), where the MMT's battery level is low and immediate access to an energy source is not possible (e.g., user not at home or office). In this case the TFAP concerns power consumption optimization and economic cost is ignored. This problem will be referred to as TFAP$_P$.

3. Any other situation, where the value of $P_{\max}$ may range in an interval $[P_a, P_b]$, where P_a, P_b represent the minimum and maximum power consumption for serving user traffic at a given networking context and user device configuration.

The execution of the TFAP solution procedure is triggered whenever the MMT perceives events related to changes in the problem parameters $\mathbf{R}$, $\mathbf{B}$, $\mathbf{F}$ or $P_{\max}$. Typical events are: (a) availability of new bearer services or imminent unavailability of already used services, (b) degradation of the QoS of one or more currently used bearer services, (c) arrival or termination of application traffic flows, (d) changes in the value of $P_{\max}$.

The determination of the current value of $P_{\max}$ is a subproblem that will be delegated to the MMT's Power Management Subsystem (PMS). Specifically, PMS will periodically evaluate and update the value of $P_{\max}$ with the objective of providing energy sufficiency to the MMT for a specified time duration D. The minimum duration of MMT energy autonomy D is a user input that represents an estimation of the time until next battery recharge. In addition to D, other factors that determine the value of $P_{\max}$ are: (a) current battery energy availability E, (b) last time instance t_D of user update of D, (c) amount of served traffic since t_D, (d) expected traffic load until the expiration of D, etc. The specification of power management policies or decision mechanisms for fixing $P_{\max}$ is out of scope of this work.

Given the value of $P_{\max}$, TFAP is a combinatorial optimization problem and methods relevant to this problem category may be applied for its solution. An integer linear programming formulation of the TFAP is presented in [ZG08]. On the basis of this formulation exact solutions of the TFAP can be produced with the help of proprietary or open source software. Such software is used for the generation of exact solutions in the experiments described in Section 3.4. However, finding exact solutions even for problems of moderate size, often requires several seconds of execution in a standard workstation PC. As a result, the execution of an exact algorithm in processing power limited mobile devices will be certainly inefficient in terms of responsiveness and utilization of device resources. On the other hand, its deployment to the network side and execution for large numbers of MMT's may raise scalability issues. For these reasons, an algorithm that establishes a good trade off between solutions' quality and responsiveness is required for the TFAP. Towards this direction this thesis introduces a heuristic algorithm for solving the TFAP that is based on local search [RN03]. The algorithm performs a guided search in the problem state space (valid flow assignments) in search of efficient solutions and is characterized by low execution times as it enumerates a relatively small subset of the actual problem state space.

3.2.3 A Binary Integer Programming formulation for TFAP

In order to produce exact solutions with off-the-self optimization software a formulation of the TFAP problem as a binary integer linear programming problem has been developed. The problem formulation is based on the identification of variables and constraints that represent the problem concepts as analyzed in Section 3.2. An initial version of this problem formulation has been presented in paper [ZG08], where the TFAP problem was initially described.

Variable definitions

Recall from the previous subsection that $\mathbf{R} = \{r_i : 1 \leq i \leq m, m \in \mathbb{N}^*\}$ represents the set of a MMT's radio interfaces. Moreover, $\mathbf{B} = \{b_j : 1 \leq j \leq n, n \in \mathbb{N}^*\}$ is the set of bearer services provided by the RANs that serve the MMT's current location. Each $b_j \in \mathbf{B}$ is characterized by a set of service, cost and power consumption attributes, i.e. $b_j = (k_{u,j}, k_{d,j}, Q_j, c_j, wr_j, wt_j, wb_j)$. The set $\mathbf{B}$ is partitioned into $m = |\mathbf{R}|$ disjoint subsets, i.e. $\mathbf{B} = B_1 \cup B_2 \ldots \cup B_m$, where each subset B_i comprises services that are compatible with radio interface r_i. Let $n_i = |B_i|$ be the cardinality of each subset B_i. For the sake of clarity the jth element of a subset B_i will be referred to as $b_{ij} = (k_{u,ij}, k_{d,ij}, Q_{ij}, c_{ij}, wr_{ij}, wt_{ij}, wb_{ij})$.

Application traffic that needs to be served by the MMT is represented by the set of traffic flows $\mathbf{F} = \{f_z : 1 \leq z \leq v, v \in \mathbb{N}^*\}$. Each flow is characterized by its bandwidth and QoS requirements, i.e. $f_z = (bw_{u,z}, bw_{d,z}, q_z)$.

The TFAP problem involves association of each $r_i \in \mathbf{R}$ with at most one $b_{ij} \in B_i$ and assignment of application traffic flows to activated radio interfaces in a way that economic cost C_M is minimized without violating problem constraints.

The Binary Integer Programming (BIP) formulation of TFAP is based on two sets of binary variables, i.e. variables that are allowed to take a value of 0 or 1. Let y_{ij} be a variable that represents the association of r_i with $b_{ij} \in B_i$:

$$y_{ij} = \begin{cases} 1, & r_i \text{ associated with } b_{ij} \in B_i \\ 0, & r_i \text{ not associated with } b_{ij} \in B_i \end{cases} \tag{3.6}$$

Let x_{ijz} be a binary variable that represents the assignment of flow f_z to radio interface r_i associated with bearer service $b_{ij} \in B_i$. It holds that:

$$x_{ijz} = \begin{cases} 1, & f_z \text{ assigned to } b_{ij} \in B_i \\ 0, & f_z \text{ not assigned to } b_{ij} \in B_i \end{cases} \tag{3.7}$$

In the rest of this section the $\mathbf{X}$ notation will be used for representing the set of x_{ijz} variables, while $\mathbf{Y}$ will refers to the set of y_{ij} variables.

Problem formulation

On the basis of the above variable definitions the objective function of TFAP is:

$$C_M(\mathbf{X}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{z=1}^{v} c_{ij}(bw_{u,z} + bw_{d,z}) x_{ijz}. \tag{3.8}$$

The TFAP problem involves minimization of $C_M(\mathbf{X})$ subject to the following constraint sets (CS):

CS_1. A radio interface r_i may be associated with at most one bearer service $b_{ij} \in B_i$. An interface that is not associated with any bearer service is assumed to be inactive (powered off).

$$\sum_{j=1}^{n_i} y_{ij} \leq 1, \forall i \in \{1, \ldots m\}. \tag{3.9}$$

CS_2. A flow f_z may be served by at most one b_{ij}. Assignment of f_z to bearer service b_{ij} also assumes that radio interface r_i is activated and associated with b_{ij}. This is guaranteed by constraint sets CS_3 and CS_4 that are specified below:

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ijz} = 1, \forall z \in \{1, \ldots v\}. \tag{3.10}$$

CS_3. Uplink capacity of each bearer service b_{ij} must not be exceeded. Note that this constraint ensures that there will be no i, j, z for which $x_{ijz} = 1$ and $y_{ij} = 0$ at the same time, i.e. f_z may not be assigned to a bearer service not associated with a MMT's radio interface.

$$\sum_{z=1}^{v} x_{ijz} bw_{u,z} - k_{u,ij}\ y_{ij} \leq 0, \tag{3.11}$$

$$\forall i, j : i \in \{1, \ldots m\}, j \in \{1, \ldots n_i\}.$$

CS_4. Downlink capacity of each bearer service b_{ij} must not be exceeded:

$$\sum_{z=1}^{v} x_{ijz} bw_{d,z} - k_{d,ij}\ y_{ij} \leq 0, \qquad (3.12)$$
$$\forall i,j : i \in \{1,\dots m\}, j \in \{1,\dots n_i\}.$$

CS_5. The QoS requirements of a flow f_z must not be violated by the provided QoS of service b_{ij} that is assigned to:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ijz} Q_{ij} - q_z \leq 0, \forall z \in \{1,\dots v\}. \qquad (3.13)$$

CS_6. The power consumption of the solution must not exceed a limit $P_{\max}$. The power consumption C_P depends on the values of variables $\mathbf{X}$, $\mathbf{Y}$ and thus, will be represented as $C_P(\mathbf{X},\mathbf{Y})$.

$$\begin{aligned} C_p(\mathbf{X},\mathbf{Y}) \leq & P_{\max}, & (3.14)\\ C_p(\mathbf{X},\mathbf{Y}) = & \sum_{i=1}^{m}\sum_{j=1}^{n_i} wr_{ij} \sum_{z=1}^{v} bw_{u,z}\ x_{ijz} & (3.15)\\ & + \sum_{i=1}^{m}\sum_{j=1}^{n_i} wt_{ij} \sum_{z=1}^{v} bw_{d,z}\ x_{ijz} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{n_i} wb_{ij}\ y_{ij}. \end{aligned}$$

3.2.4 Problem complexity

The maximization version of TFAP is a generalization of the Multiple Knapsack problem with Assignment Restrictions (MKAR). MKAR is a variant of the Multiple Knapsack Problem (MKP) [MT90] and is NP-Hard [DKK$^+$00]. MKAR is defined as the problem of finding a maximum profit assignment of n items $u_j \in U$ to m knapsacks $v_i \in V$ of finite capacity. Each item u_j has a weight w_j, independent of the knapsack that is assigned to, and its assignment to any knapsack brings in profit p_j. Moreover, each knapsack v_i has a finite capacity k_i and sets restrictions on the items that it can admit. Let $N_i \subseteq U$ be the set of items that are allowed to be assigned to a knapsack v_i. The MKAR problem involves finding a partition of U into m disjoint subsets $U = U_1 \cup U_2 \ldots \cup U_m, U_i \subseteq N_i, \forall i \in \{1, ..m\}$, for assignment of their items to corresponding knapsacks v_i, in a way that maximizes profit without violating knapsack capacity constraints:

$$\begin{aligned} &\text{maximize} && M = \sum_{v_i \in V} \sum_{u_j \in U_i} p_j \\ &\text{subject to} && \sum_{u_j \in U_i} w_j \leq k_i, && \forall v_i \in V. \end{aligned}$$

MKAR is NP-Hard in the strong sense even in the case of equal knapsack capacities [DKK$^+$00]. TFAP is at least as difficult as MKAR which can be proved by restriction [GJ90], i.e., by showing that TFAP contains the MKAR as a special case. However, MKAR is a maximization problem and, thus, the proof will be based on the maximization version of TFAP, *MAX-TFAP*. MAX-TFAP is equivalent to TFAP, i.e., a MAX-TFAP solution is also a solution to TFAP and vice versa, and derives from TFAP through a simple transformation.

Assume a TFAP problem instance $(\mathbf{R}, \mathbf{B}, \mathbf{F}, P_{\max})$ and let t be a positive real number with $t > \max_{b_j \in \mathbf{B}} c_j$. The profit associated with the use of a bearer service b_j is $p_j = t - c_j$ and has the same measurement unit with c_j. Moreover, let W be the total traffic requirements in kbps of the TFAP instance, i.e., $W = \sum_{f_z \in \mathbf{F}} (bw_{u,z} + bw_{d,z})$. The assignment of a flow f_z to a service b_j yields profit $p_{jz} = p_j(bw_{u,z} + bw_{d,z})$, while the total profit of a traffic flow assignment S, that has a corresponding radio interface activation B', is given by function $P_M(S)$:

$$\begin{aligned}
P_M(S) &= \sum_{b_j \in B'} p_j \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}) \\
&= \sum_{b_j \in B'} (t - c_j) \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}) \\
&= t \sum_{b_j \in B'} \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}) - \sum_{b_j \in B'} c_j \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}) \\
&= tW - C_M(S).
\end{aligned}$$

Note that $\sum_{b_j \in B'} \sum_{f_z \in F_j} (bw_{u,z} + bw_{d,z}) = W$, since by the definition of a traffic flow assignment the union of all subsets F_j is equal to $\mathbf{F}$. The first term of $P_M(S)$ is constant for a given TFAP instance and independent of S, thus, a flow assignment that minimizes $C_M(S)$ maximizes $P_M(S)$ and vice versa. MAX-TFAP involves maximization of $P_M(S)$ subject to the same constraints with TFAP. Therefore, a solution of a MAX-TFAP problem instance is also a solution of its corresponding TFAP instance.

In this thesis, the proof on the hardness of MAX-TFAP involves the specification of restrictions on its instances so that the resulting problem instances are identical to MKAR [GJ90]. Let ($\mathbf{R}$, $\mathbf{B}$, $\mathbf{F}$) be a MAX-TFAP instance with $P_{\max}$ limit on power consumption. Recall that $\mathbf{B} = B_1 \cup B_2 \cup \ldots \cup B_m$, where each subset B_i includes services compatible with a specific radio interface r_i. The following restrictions are applied to the elements of each $B_i = \{b_1, b_2, \ldots b_{t_i}\}$:

- $k_{u,1} = k_{u,2} \ldots = k_{u,t_i}$, i.e., uplink capacities of bearer services are equal to each other,
- $k_{d,1} = k_{d,2} \ldots = k_{d,t_i}$, i.e., downlink capacities of bearer services are equal to each other,
- $Q_1 = Q_2 = \ldots = Q_{t_i}$, i.e., provided QoS is equal for all bearer services in B_i,
- $wb_1 = wb_2 = \ldots = wb_{t_i}, wr_1 = wr_2 = \ldots = wr_{t_i}, wt_1 = wt_2 = \ldots = wt_{t_i}$, i.e., power consumption coefficients are equal to each other.

Moreover, assume that all bearer services have the same cost, i.e., $c_j = c, \forall b_j \in \mathbf{B}$, and $t = c + 1$. Then, the profit p_j corresponding to each service b_j is $p_j = t - c = 1$. Regarding traffic flows, it is assumed that they are all downlink, i.e., $bw_{u,z} = 0, \forall f_z \in \mathbf{F}$.

Since bearer services that are included in each subset B_i are identical in terms of their service, cost and power consumption attributes, the restricted problem does not involve bearer selection at a radio interface level. Thus, each restricted problem instance is based on a set $B'' \subseteq \mathbf{B}, |B''| = m$, that includes one randomly selected element from each subset $B_i, 1 \leq i \leq m$. The restricted problem involves finding an assignment of flows to bearer services in B'' (i.e., a partition of $\mathbf{F}$ into m disjoint sets) that maximizes profit subject to bearer services' capacity constraints and traffic flows' QoS requirements. The existence of bearer services without any assigned flows in a problem solution implies that the respective radio interfaces are deactivated.

The restriction that is applied to the values of $P_{\max}$ enables the simultaneous use of all m radio interfaces for serving application traffic without exceeding the maximum allowed power consumption: $P_{\max} = \sum_{b_j \in B''} (wb_j + wr_j k_{d,j})$. Thus, the constraint on power consumption can be eliminated from the restricted version of MAX-TFAP.

With reference to the MKAR problem definition, the profit from assigning a flow f_z to any bearer service b_j is $p_z = p_j \; bw_{d,z} = bw_{d,z}$, while the "weight" of each flow is $w_z = bw_{d,z}$ that is also independent of the service that is assigned to. Let N_j be the subset of flows that are allowed to be assigned to a service b_j on the basis of their QoS requirements, i.e., $N_j = \{f_z \in \mathbf{F} : q_z \leq Q_j\}$. The constraint represented by equation (3) in the formulation of TFAP can, then, be rewritten as $F_j \subseteq N_j$.

On the basis of the aforementioned assumptions and restrictions the restricted MAX-TFAP problem can be stated as: Find a partition of $\mathbf{F}$ into $m = |B''|$ disjoint subsets $F_j, j \in \{1, \ldots m\}$, each one corresponding to a bearer service $b_j \in B''$, that maximizes profit without violating problem constraints:

$$
\begin{array}{lll}
\text{maximize} & M = \displaystyle\sum_{b_j \in B''} \sum_{f_z \in F_j} p_z & \\
\text{subject to} & \displaystyle\sum_{f_z \in F_j} w_z \leq k_{d,j}, & \forall b_j \in B'', \\
 & F_j \subseteq N_j, & \forall b_j \in B''.
\end{array}
$$

This problem formulation corresponds to the MKAR optimization problem and, thus, MAX-TFAP and its equivalent TFAP are NP-hard.

3.3 A Heuristic Algorithm for the TFAP

In this section a heuristic algorithm is introduced for approximating the solution of a TFAP instance. It represents an iterative improvement algorithm that performs an exploration in the space of problem's feasible solutions (or problem states) in search of the minimum cost solution [RN03]. The cost of each problem state is evaluated with an appropriate objective function. The algorithm starts from initial problem states and iteratively modifies them in order to reach lower cost solutions that constitute better approximations of the global solution. Specifically, given any initial problem state, the algorithm performs local search, i.e., it follows a path of neighbouring states in a way that the objective function value is decreasing across the path. Thus, the algorithm belongs to the category of "hill climbing" search algorithms [RN03]. Initial problem solutions are obtained through appropriate construction algorithms. These algorithms "construct" a valid traffic flow assignment S by assigning flows one-by-one to appropriate radio interfaces with a first-fit assignment scheme. Assuming an ordered set of radio interfaces, first-fit assignment places each flow to the first activated interface that satisfies its bandwidth and QoS requirements.

A known weakness of hill climbing algorithms is that they often get stuck to local minima, thus failing to reach a global solution. This issue is often handled with random-restarts, i.e., with successive searches from randomly generated initial states. Another approach that focuses on overcoming local minima is the "simulated annealing" algorithm that combines the hill climbing search with a random walk [RN03]. The proposed heuristic algorithm employs a limited number of random restarts due to requirements of the TFAP domain for frequent and efficient real-time algorithm execution.

Algorithm 1 gives an overview of the TFAP solution procedure and presents the synergy between construction heuristics and search. The notation S_x represents a valid traffic flow assignment. In the rest of this section the terms problem state and problem solution will also be used, denoting a valid traffic flow assignment as it is specified in Definition 2 of Section 3.2.1. Procedures `findMinimumCostSolution` and `findMinimumPwrConsSolution` implement the First Fit construction heuristics that generate the initial problem states (S_M, S_P) for local search. Initial problem states are constructed with focus on the optimization of a single objective. Thus, S_M approximates the minimum economic

Algorithm 1: Heuristic Traffic Flow Assignment Algorithm

```
input  : Power consumption limit $P_{\max}$, Radio interfaces $R$, Bearer services $B$, Flows $F$
output: A valid traffic flow assignment $S$
$S_M \leftarrow$ findMinimumCostSolution($R$,$B$,$F$);
$S_P \leftarrow$ findMinimumPwrConsSolution($R$,$B$,$F$);
$S_1 \leftarrow$ localSearchRS($S_M$, $P_{\max}$);
$S_2 \leftarrow$ localSearchRS($S_P$, $P_{\max}$);
/* If $S_1$ has lower economic cost and respects the limit $P_{\max}$ */
if $C_M$($S_1$) < $C_M$($S_2$) and $C_P$($S_1$) $\leq P_{\max}$ then return $S_1$;
/* If $S_2$ has lower economic cost and respects the limit $P_{\max}$ */
if $C_M$($S_2$) < $C_M$($S_1$) and $C_P$($S_2$) $\leq P_{\max}$ then return $S_2$;
/* If neither of the above was true then it holds that either $C_M$($S_1$) =
   $C_M$($S_2$) or the limit $P_{\max}$ cannot be met.  Select the lowest power
   consumption solution. */
if $C_P$($S_1$) < $C_P$($S_2$) then return $S_1$;
else return $S_2$;
```

cost problem solution, while S_P approximates the minimum power consumption one. The functions $C_M(S)$ and $C_P(S)$ return the economic and power consumption cost of a given problem state S. Procedure `localSearchRS` implements the local search with restarts algorithm and is described in Section 3.3.2. The local search part of the algorithm is implemented by procedure `localSearch` that is specified in Section 3.3.1.

The notation that will be used for the description of the algorithms corresponds to that introduced in Sections 3.2.1 and 3.2.4. Regarding the bearer service concept, additional notation will be introduced for a more concise description of the algorithms. Recall from Section 3.2.4 that $N_j \subseteq \mathbf{F}$ represents the set of "admissible" traffic flows for a bearer service b_j, i.e., the set of flows that b_j satisfies their QoS requirements and, thus, can be assigned to it (when associated with an appropriate radio interface). The "effective" uplink and downlink bandwidth capacity of a bearer service b_j are defined, respectively, as:

$$k'_{u,j} = \min\,(k_{u,j}, \sum_{f_z \in N_j} bw_{u,j})$$

$$k'_{d,j} = \min\,(k_{d,j}, \sum_{f_z \in N_j} bw_{d,j}).$$

Based on the notion of effective capacity, the *power consumption coefficient* w_j is introduced for a bearer service b_j, that represents the power consumption per unit of effective capacity of the bearer service:

$$w_j = \frac{wb_j + wr_j \; k'_{d,j} + wt_j \; k'_{u,j}}{k'_{d,j} + k'_{u,j}}. \tag{3.16}$$

$$\tag{3.17}$$

In the rest of this section a distinction will be made among the power consumption of a radio interface r_i due to its currently assigned flows, *total power consumption*, and the coefficient w_j that refers to its currently associated bearer service b_j. Accordingly, the *total economic cost* of a radio interface refers to the cost due to its currently assigned flows in a given problem state S, while the cost c_j corresponds to the cost of its associated bearer service b_j.

3.3.1 Local search algorithm

The hill climbing algorithm is represented by procedure `localSearch` that accepts as input parameters an initial problem state and an upper limit on power consumption $P_{\max}$. Given the initial problem state, the procedure performs successive transitions to other problem states with aim to converge to one that has the minimum possible economic cost and a power consumption that does not violate the upper limit $P_{\max}$. A series of state transitions that decrease C_M are usually followed by an increase in C_P. This is due to the gradual activation of available radio interfaces, as traffic flows are distributed to the cheapest possible services that are accessible through each radio interface. The assumption here is that traffic requirements exceed the capacity of the cheapest bearer service and bearer service cost is positively correlated with its QoS and capacity. On the other hand, state transitions that gradually decrease C_P probably increase C_M, given that higher capacity, higher QoS bearers usually have higher usage cost. This potential trade-off between economic cost and power consumption is taken into account in the design of the local search strategy. Thus, local search is guided towards either: (a) optimization of C_M with minimum increase in C_P (C_M minimization search iteration), if $C_P(S_i) \leq P_{\max}$ holds for the initial state S_i or (b) optimization of C_P with minimum increase in C_M (C_P minimization search iteration), if $C_P(S_i) > P_{\max}$. A graphical representation of the `localSearch` procedure, with emphasis on the interchange of search iterations, is depicted in Figure 3.1.

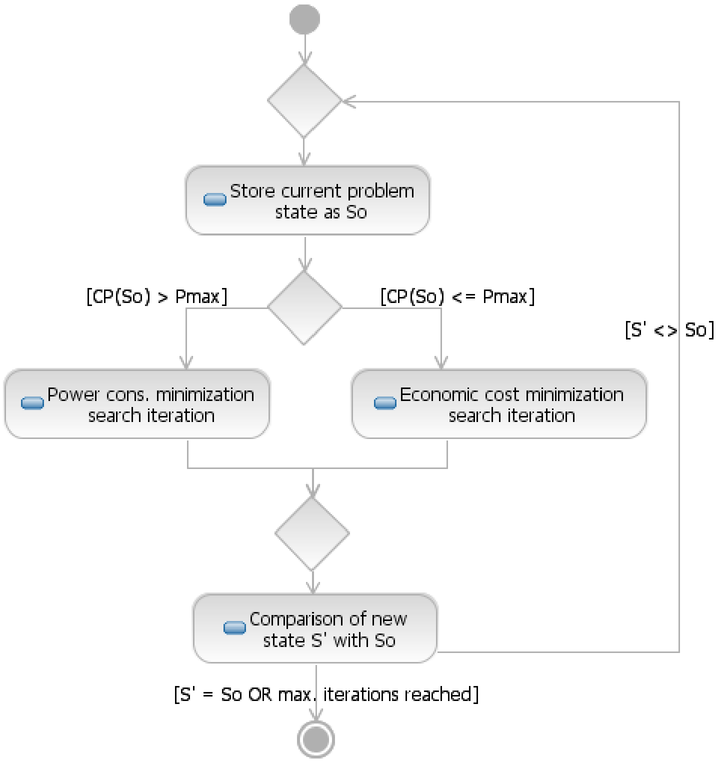
Store current problem state as So
[CP(So) > Pmax]
[CP(So) <= Pmax]
[S' <> So]
Power cons. minimization search iteration
Economic cost minimization search iteration
Comparison of new state S' with So
[S' = So OR max. iterations reached]

Figure 3.1: Overview of the local search algorithm.

The core part of the `localSearch` procedure is described in Algorithm 2. It represents a local search iteration that targets economic cost optimization. Thus, the condition $C_P(S) \leq P_{\max}$ must be true for the algorithm parameter S (initial problem state). If $C_P(S) > P_{\max}$ then algorithm execution will not lead to a state transition. However, in this case the control logic of the `localSearch` procedure will invoke a variation of Algorithm 2 that performs power consumption optimization until reaching the limit $P_{\max}$. Its differences from Algorithm 2 are localized in statements 1, 8, and 11 and their variants are included as comments in the algorithm. Specifically, (a) in statement 1 radio interfaces are sorted by descending w_j of their associated bearers services first and then by *total economic cost*, (b) in statement 8 the last parameter of `searchStep` is set to *false* and (c) in statement 11 a different objective function is used (function g). The role of the boolean parameter in `searchStep` will be explained thereafter. The control flow of `localSearch` involves the iterative execution of Algorithm 2 and its variation. The value of $C_P(S')$, where S' is the problem state after a local search iteration, determines which of the two algorithms will be executed in the next iteration. Their execution is, thus, interchanged until a maximum number of iterations is reached or until two successive iterations result in identical problem states.

Algorithm 2: Local search iteration for cost minimization

```
input  : Initial problem state $S$, power consumption limit $P_{\max}$
output: A new problem state $S'$
Sort radio interfaces $R$ by descending cost $c_j$, total power consumption ;
/*  Sort radio interfaces $R$ by descending $w_j$, total economic cost      */
foreach radio interface $r$ in $R$ do
    $F_r \leftarrow$ servedFlows($r$) ;
    Sort $F_r$ by descending QoS $q_z$, ascending bandwidth $(bw_{u,z} + bw_{d,z})$ ;
    $R' \leftarrow R \setminus \{r\}$ ;
    foreach flow $f_z$ in $F_r$ do
        foreach $r_i \in R', i \in \{1, .., |R'|\}$ do
            $S_i \leftarrow$ searchStep($r$, $f_z$, $r_i$, true) ;
            /*  $S_i \leftarrow$ searchStep($r$, $f_z$, $r_i$, false)              */
            Restore state $S$ ;
        end
        /*  Select a state transition that minimizes the value of f.    */
        $j \leftarrow \arg_i \min$ f($S$, $S_i$, $P_{\max}$) ;
        /*  $j \leftarrow \arg_i \min$ g($S$, $S_i$, $P_{\max}$)                    */
        Restore state $S_j$ ;
        /*  Continue and set as initial state the $S_j$.                 */
        $S \leftarrow S_j$ ;
    end
end
/*  Variable S now holds the final state after all applied transitions.
    */
return $S$
```

The main idea behind the algorithm for the local search iteration is the permutation of flows across pairs of radio interfaces in a way that the objective function is minimized. Specifically, starting from the most expensive or energy demanding radio interface (depending on the algorithm variation), a permutation procedure is executed for each served flow against the other radio interfaces. This permutation procedure, called `searchStep`, is described by Algorithm 3 and results to: (a) no state transition or (b) transition to a new valid problem state. Thus, `searchStep` constitutes the neighborhood operation of the local search algorithm. The possible state transitions concerning a certain flow f_z are evaluated by an objective function and the algorithm continues from the state that minimizes its value.

The neighborhood operation `searchStep` requires four parameters: (a) a source radio interface r_s, (b) a traffic flow f, served by r_s, (c) a target radio interface r_t and (d) a boolean `optimizeCM` that is set to true, when `searchStep` is invoked during an economic cost optimization iteration, or false when invoked during power consumption minimization. Its purpose is to assign f to r_t by associating r_t with an appropriate bearer and possibly replacing one or more of its served flows. The function `candidateBearers(r)` returns the bearer services of the same RAT with radio interface r, while the function `associatedBearer(r)` returns the bearer service that is associated with r in the current problem state. Procedure `searchStep` tries to assign f to r_t by checking each candidate bearer service, starting from b_{ta} and continuing with ascending cost c_j or ascending w_j, depending on the value of `optimizeCM`.

Algorithm 3: Procedure searchStep

```
input  : Radio interface $r_s$, Flow f, Radio interface $r_t$, boolean optimizeCM
output: A new traffic flow assignment $S$
$B_t \leftarrow$ candidateBearers($r_t$) ;
/*  Get currently associated bearers with $r_t$, $r_s$                      */
$b_{ta} \leftarrow$ associatedBearer($r_t$ ) ;
$b_{sa} \leftarrow$ associatedBearer($r_s$ ) ;
if optimizeCM then
    Sort $B_t$ by ascending $c_j$, $w_j$, descending $Q_j$ and $(k_{d,j} + k_{u,j})$ ;
else
    Sort $B_t$ by ascending $w_j$, $c_j$, descending $Q_j$ and $(k_{d,j} + k_{u,j})$ ;
end
foreach bearer $b_j$ in $B_t$ starting from $b_{ta}$ do
    $F_t \leftarrow$ append($f$, $b_j$) ;
    if $F_t \neq \{f\}$ then
        /*  $f$ was successfully assigned to $b_j$ $(F_t = \emptyset)$ or $f$ was assigned to
            $b_j$ and replaced other flows                                   */
        break;
    end
    /*  else try to assign $f$ to the next bearer                           */
end
/* Assign displaced flows to $r_s$                                           */
foreach flow $f_t$ in $F_t$ do
    append($f_t$, $b_{sa}$) ;
end
return current flow assignment $S$
```

Flow assignment is performed by procedure `append` that (i) assigns f to b_j, if it has spare bandwidth (note that b_j is considered to be serving the flows already assigned to its corresponding radio interface r_t), or (ii) replaces one or more served flows of r_t so as to save the required bandwidth. Flow replacement applies to flows that have QoS requirements equal or lower than f and their aggregate bandwidth requirements are lower than those of f. Thus, `append` returns a set of flows F_t that contains: (a) $\emptyset$ if f was assigned to b_j without replacing any flows, (b) $\{f\}$ if assignment of f was not possible, (c) one or more served flows of r_t. Replaced flows F_t are assigned to r_s that has the required capacity, as replaced flows have an aggregate bandwidth that is less than the required bandwidth of f.

Equations 3.18 and 3.19 describe the objective functions f and g, that are used respectively in Algorithm 2 and its variation for power consumption minimization. Each objective function evaluates a state transition from state S_a to state S_b given a power consumption limit P. Function f favors state transitions that reduce economic cost and also cause the least increase in power consumption per unit of cost decrease. State transitions that either violate the power consumption limit or increase the economic cost are mapped to a very large constant K. Finally, no state transition ($S_a \equiv S_b$) results to a function value of 0. Objective function g has a similar role that favor's power consumption minimization.

$$f(S_a, S_b, P) = \begin{cases} \frac{C_P(S_b) - C_P(S_a)}{|C_M(S_b) - C_M(S_a)|}, & C_P(S_b) \leq P \wedge C_M(S_b) < C_M(S_a) \\ 0, & S_a \equiv S_b \\ K \to \infty, & \text{in any other case,} \end{cases} \tag{3.18}$$

$$g(S_a, S_b, P) = \begin{cases} \frac{C_M(S_b) - C_M(S_a)}{|C_P(S_b) - C_P(S_a)|}, & C_P(S_b) \leq P \wedge C_P(S_b) < C_P(S_a) \\ 0, & S_a \equiv S_b \\ K \to \infty, & \text{in any other case.} \end{cases} \tag{3.19}$$

Functions f' and g' (equations 3.20, 3.21) represent objective functions, alternatives to f and g respectively, that may also be applied for the evaluation of state transitions. The employment of these functions results to more aggressive minimization of the respective objectives during local search. The reason is that each function favours state transitions with maximum cost decrease, either economic cost or power consumption, rather than transitions that establish the best trade-off between the two objectives. The functions will be henceforth referred to as "greedy" objective functions. However, as it is shown during algorithm's approximation performance evaluation (Section 3.4.1), f' and g' result to inferior problem solutions.

$$f'(S_a, S_b, P) = \begin{cases} C_M(S_b) - C_M(S_a), & C_P(S_b) \leq P \wedge C_M(S_b) < C_M(S_a) \\ 0, & S_a \equiv S_b \\ K \to \infty, & \text{in any other case,} \end{cases} \tag{3.20}$$

$$g'(S_a, S_b, P) = \begin{cases} C_P(S_b) - C_P(S_a), & C_P(S_b) \leq P \wedge C_P(S_b) < C_P(S_a) \\ 0, & S_a \equiv S_b \\ K \to \infty, & \text{in any other case.} \end{cases} \tag{3.21}$$

Algorithm 4: Procedure findMinimumCostSolution

```
input  : Radio interfaces $R$, Bearer services $B$, Flows $F$
output: An assignment $S$ of bearer services and flows to radio interfaces
/*  First Fit flow assignment heuristic                                    */
Sort $B$ by ascending cost $c_j$, descending bandwidth $(k_{u,j} + k_{d,j})$ ;
/*  Sort $B$ by ascending $w_j$, descending bandwidth $(k_{u,j} + k_{d,j})$  */
Sort $F$ by descending bandwidth $b_z$, descending QoS $q_z$ ;
foreach flow $f_z$ in $F$ do
    $F_0 \leftarrow \{f_z\}$ ;
    foreach bearer $b_j$ in $B$ do
        $F' \leftarrow \emptyset$;
        foreach flow $f_k$ in $F_0$ do
            /*  Assign $f_k$ to $b_j$ and, if required, replace 1 or more flows
                currently served by the bearer's respective radio
                interface.  Displaced flows will be assigned to subsequent
                bearers.                                                   */
            $F' \leftarrow F' \cup$ append($f_k$, $b_j$);
        end
        $F_0 \leftarrow F_0 \cup F'$;
        /*  Flow $f_z$ and flows displaced by it have been assigned to a
            bearer.  Proceed with next flow in $F$.                        */
        if $F_0 = \emptyset$ then break;
    end
end
return current flow assignment $S$
```

Finally, Algorithm 4 describes the construction of problem solutions that will be used as initial states in local search. The algorithm describes the procedure `findMinimumCostSolution` of Algorithm 1, that approximates the minimum cost solution S_M of a TFAP instance. The algorithm performs a first fit assignment of flows to bearer services with priority to low cost services. Assignment takes place with the use of `append` that was also used in `searchStep`. The construction of S_M is based on the principle of reserving capacity (through assignment of flows) from the most cost efficient bearers. Thus, bearer services are sorted in ascending cost order prior to flow assignment. In case that two or more bearers with the same cost have different capacity, priority is given to the highest capacity bearer. The idea is to utilize as much as possible capacity at a given cost and leave less traffic to be assigned to subsequent and, thus, more expensive bearers.

The minimum power consumption solution S_P of a problem instance is approximated by a variation of Algorithm 4 that differs from it in statement 1, where bearer services are sorted by ascending w_j and then by descending bandwidth. In this algorithm version the power consumption represents the cost factor and, thus, priority is given to capacity reservation on the most power efficient bearers. The variant of this statement is included as a comment in the algorithm specification.

3.3.2 Handling local minima

The hill climbing local search algorithm, implemented by procedure `localSearch`, returns at the end of its execution the minimum C_M problem state, encountered during its search, that does not violate the upper limit $P_{\max}$ in power consumption. However, the returned problem state does not necessarily represent a global minimum and, consequently, an exact solution of the TFAP problem instance. Thus, in the proposed TFAP heuristic algorithm, the outcome of each `localSearch` execution is treated as a local minimum and a random restart of the local search is initialized. However, for reasons of execution efficiency, the restart of the local search takes place for a limited number of iterations.

Algorithm 5: Procedure localSearchRS

```
input  : Power consumption limit $P_{\max}$, Problem state $S_o$
output: A valid traffic flow assignment $S$
$S_r \leftarrow$ localSearch($S_o$, $P_{\max}$ ) ;
foreach Radio interface $r_i$ in $R$ do
    $S_i \leftarrow$ findRestartState($S_r$, $r_i$ ) ;
    $S_r \leftarrow$ localSearch($S_i$, $P_{\max}$ ) ;
end
$S \leftarrow S_r$ ;
return $S$;
```

Algorithm 5 specifies the `localSearchRS` procedure, that combines local search with random restarts through invocation of procedures `localSearch` and `findRestartState`.

Specifically, given an initial solution, constructed through Algorithm 4, `localSearchRS` executes local search with a series of restarts for each radio interface. Procedure `findRestartState` is responsible for the generation of the initial state S_i for each restart of the local search algorithm.

Algorithm 6: Procedure findRestartState

```
input  : Initial problem state $S$ , Radio interface $r_t$
output: A new problem state $S'$
$b_{ta} \leftarrow$ associatedBearer($r_t$ ) ;
Randomly select a QoS value $Q_0 < Q_{ta}$ ;
$F_r \leftarrow$ servedFlows($r_t$ ) ;
/* Served flows of $r_t$ with higher QoS than $Q_0$.                  */
$F_0 \leftarrow \{f_z \in F_r : q_z > Q_0\}$ ;
Remove $F_0$ from the served flows of $r_t$;
/* Compatible bearer services of $r_t$ with higher QoS than $Q_0$.    */
$B_0 \leftarrow \{b_j \in B_t : Q_j > Q_0\}$ ;
/* Remove the set $B_0$ from compatible bearers of $r_t$.             */
$B_t \leftarrow B_t \setminus B_0$ ;
$R' \leftarrow R \setminus \{r_t\}$ ;
foreach flow $f_z$ in $F_0$ do
    foreach $r_i \in R', i \in \{1, .., |R'|\}$ do
        $S_i \leftarrow$ searchStep($r_t$, $f_z$, $r_i$, true) ;
        Restore state $S$ ;
    end
    /*  Select a state transition that minimizes the cost increase.   */
    $j \leftarrow \arg_i \min(C_M(S_i) - C_M(S))$ ;
    Restore state $S_j$ ;
    /*  Continue and set as initial state the $S_j$.                  */
    $S \leftarrow S_j$ ;
end
/* Add $B_0$ to the set of compatible bearers of $r_t$.               */
$B_t \leftarrow B_t \cup B_0$ ;
/*  Variable S now holds the final state after all applied transitions.
    */
return $S$ ;
```

Algorithm 6 specifies the `findRestartState` procedure. It admits two parameters, (a) a problem state S and (b) a radio interface r_t, and returns a new problem state S' for local search restart. The restart state is produced by constraining r_t to use a bearer service of a lower, randomly selected, QoS class than the one it was associated to at the initial state S (bearer b_{ta} in Algorithm 6). The main idea is to remove flows with high QoS requirements from the radio interface r_t, and, thus, to allow the local search to evaluate bearer services of lower QoS that, although more expensive, provide higher capacity than b_{ta}. The removed flows are assigned to other radio interfaces, with a minimum cost increase, by executing part of the local search iteration algorithm (Algorithm 2).

3.4 Evaluation of the heuristic algorithm

3.4.1 Solution accuracy and runtime performance

Experimental setting

The evaluation of the local search algorithm for the TFAP is based on the comparison of heuristic and exact solutions for a large number of randomly generated problem instances. Heuristic solutions are produced from a Java implementation of the proposed algorithm for the TFAP, while exact solutions result from a tool capable of solving Integer Linear Programming problems (Lingo [LIN09]) that will be henceforth referred to as *ILP solver*. The random generation of problem instances is based on four problem templates, corresponding to typical use case scenarios that motivate this work. Each problem template specifies the problem dimensions, i.e., the number of flows, radio interfaces and available bearers per radio interface. The other problem parameters are randomly generated, resulting to problem instances that are characterized by different capacity availability and bandwidth requirements.

The four use case scenarios are characterized by different types of terminal devices where the algorithm logic is deployed, and different numbers of simultaneous users. These scenarios are described below:

- UC_0, UC_1: A single mobile user is equipped with a smart-phone or netbook (MT1) that integrates two and three radio interfaces (e.g., 3GPP, 802.11a/b, WiMax) respectively. The user is engaged to various application sessions resulting to a total number of 7 flows in UC_0 and 11 traffic flows in UC_1, either inbound or outbound with different QoS and bandwidth requirements.
- UC_2: The terminal device is a notebook (MT2) that integrates 4 radio interfaces (3GPP, WiMax, 2 x IEEE 802.11a/b). The notebook serves its user's traffic, as well as traffic generated by other users working in the same team and using MT2 as an Internet gateway. It is assumed that the connectivity between MT2 and other user terminals is established through Bluetooth, forming thus a personal area network. The traffic comprises 19 flows, either inbound or outbound, with different QoS and bandwidth requirements.

Table 3.2: Flows that constitute user traffic in the use case scenarios.

Flow Id	Session type	Flow Type	Direction	QoS	Bandwidth requirements(kbps)
F1	Video conference	Audio	In	C	16,32,64
F2		Audio	Out	C	16,32,64
F3		Video	In	C	64, 128, 192, 384
F4		Video	Out	C	64, 128, 192, 384
F5	Voice Call	Audio	In	C	16,32,64
F6		Audio	Out	C	16,32,64
F7	Video Streaming	Video	In	S	64, 128, 192, 384
F8	Audio Streaming	Audio	In	S	16, 32, 64
F9	Email download	Data	In	I	64, 128, 256
F10	Email upload	Data	Out	I	64, 128, 256
F11	Ftp upload	Data	In	B	64, 128, 256, 384, 512
F12	Ftp download	Data	Out	B	64, 128, 256, 384, 512

· UC_3: A vehicular network setting is assumed in this use case scenario, where the access device is a mobile router (MT3) with 6 radio interfaces. The mobile router is incorporated in a vehicle (e.g., car or bus) and serves the traffic generated by passengers on board. The traffic comprises 32 inbound or outbound flows at various QoS levels.

The different traffic flows that are served by the terminal devices in the various use case scenarios are described in Table 3.2. The *Flow Id* column assigns an identifier to each flow for referencing purposes. Column *Session type* describes the user session that each flow is part of, while *Flow type* refers to the type of content that a flow transfers. *Direction* refers to the flow direction, either inbound or outbound, and the *QoS* column states the QoS requirements of each flow. These requirements are specified by means of a QoS class value that best matches the delay and jitter requirements of the user session. The QoS classes that are considered are documented by 3GPP and range over: (i) Conversational - C, (ii) Streaming - S, (iii) Interactive - I and (iv) Background - B. Finally, *Bandwidth requirements* column refers to alternative bandwidth requirements of a flow depending on user preferences, mobile terminal capabilities, available codecs etc. During random problem instance generation the bandwidth requirements of flows are randomly selected among the values contained in the aforementioned column.

Table 3.3 presents the served traffic in each use case scenario in terms of traffic flows

Table 3.3: Traffic load served in use case scenarios

Use Case Scenario	Active flows	Number of flows
UC_0	F1,F2,F3,F4,F9,F11,F12	7
UC_1	F1,F2,F3,F4, F7, F9, F10, 2×(F11,F12)	11
UC_2	2×(F1,F2,F3,F4), F7, 2×(F9, F10), 3×(F11,F12)	19
UC_3	3×(F1,F2,F3,F4), F5, F6, F7, F8, 3×(F9, F10), 4×(F11,F12)	32

Table 3.4: Bearer capacity availability scenarios

Capacity availability	Bandwidth capacity values (kbps)
L	64, 128, 192, 256, 320, 384
M	256, 320, 384, 448, 512
H	384, 448, 512, 576, 640, 704, 768, 832

that are defined in Table 3.2. The multiplication of a set of flows by a number denotes the number of active instances of the respective flows in the use case scenario. Thus, 2x(F11, F12) means that the mobile terminal serves two flows of type F11 and two of type F12. Concerning the radio access availability, the accessibility of 6 providers per radio interface is assumed, each one providing Internet access at four QoS level (C, S, I, B). Thus, the total available number of bearer services per radio interface is 24 (6 Conversational, 6 Streaming, 6 Interactive and 6 Background). The RAT for each one of the 24 bearer services, accessible through a specific radio interface, is selected in a random manner among the radio interface's supported RATs. The supported RATs for a 3GPP radio interface are UMTS and GPRS, while for a WLAN radio interface the supported RATs are IEEE 802.11a and IEEE 802.11b. The cost of each bearer service is randomly generated while its capacity is randomly selected from a number of predefined capacity values. Three scenarios of capacity availability are considered for all bearer services: Low (L), Medium (M) and High (H). For each capacity availability scenario, bearer capacities (equal for both uplink and downlink) are randomly selected from the value sets included in Table 3.4.

For each combination of use case and capacity availability scenario, an experiment is conducted for the evaluation of the TFAP heuristic algorithm. Thus, the proposed algorithm's evaluation process involves 12 experiments, each one corresponding to a different combination of problem size and capacity availability. The tasks that are performed during each experiment are:

1. Generation of a large number of problem instances by randomly fixing the values of: (a) flow bandwidth, (b) bearer service uplink and downlink capacity, (c) bearer service cost, (d) bearer RAT. Each problem instance is characterized by different bandwidth requirements and capacity availability.

2. Finding of the exact solution S_e and the heuristic solution S_h of each problem instance for 10 different limits on power consumption ($P_i, i \in \{1, .., 10\}$). Therefore, each problem instance produces 10 pairs of exact and heuristic solutions (S_e, S_h). The set $P = \cup\{P_i\}$ for each problem instance is generated after finding its minimum cost solution S_m and minimum power consumption solution S_p with the help of the ILP solver. Let P_m be the power consumption of S_m and P_p the respective value for S_p. As S_p corresponds to a global minimum on power consumption, it holds that $P_m \geq P_p$. The equality denotes a single global solution S_m for the multi-objective optimization problem and, thus, the set of limits contains just the value P_p. If $P_m > P_p$ then the set of limits comprises 10 equidistant values in the interval $[P_p, P_m]$, including P_p, P_m.

3. Calculation for each pair of solutions (S_e, S_h) of the approximation error $e_m = \frac{C_M(S_h) - C_M(S_e)}{C_M(S_e)}$ of the heuristic solution against the optimal one in terms of economic cost. In the following section a product of e_m by 100 will be used and referred to as *percent approximation error*. An equivalent metric e_p is also calculated for the power consumption.

Computational results

Table 3.5 summarizes the algorithm's performance results for 12 experiments that correspond to use case scenarios UC_0-UC_3. The results that are presented and analyzed correspond to 900 different problems per experiment, each on being solved for 10 different limits on power consumption. Thus, each experiment is evaluated on the basis of about 9000 exact and heuristic problem solutions.

The first column (*Exp. Id*) of table 3.5 assigns an identifier to each experiment, while column *Description* refers to the combination of use case and capacity availability scenario that characterize each experiment. The third column includes the average percent approximation error in economic cost per experiment, while the fourth presents the respective average percent approximation error in power consumption. The negative values denote

Table 3.5: Algorithm performance results per experiment

Exp. Id	Description	$\overline{e}_m$%	$\overline{e}_p$%	$\overline{T}_h$ (ms)	$\overline{T}_e$ (ms)
1	UC_0, L	4.32	-1.49	4.68	174.90
2	UC_0, M	3.99	-1.52	5.12	181.26
3	UC_0, H	4.94	-2.51	5.88	194.89
4	UC_1, L	6.30	-2.39	14.35	972.07
5	UC_1, M	5.74	-2.03	16.74	1263.35
6	UC_1, H	5.32	-2.03	22.46	1040.07
7	UC_2, L	7.39	-2.17	58.74	4702.56
8	UC_2, M	8.11	-2.36	68.50	5160.61
9	UC_2, H	5.40	-2.40	91.82	4868.94
10	UC_3, L	11.89	-3.45	400.31	9459.97
11	UC_3, M	13.07	-3.51	536.60	11157.30
12	UC_3, H	7.44	-2.95	600.06	16442.40

that the power consumption of heuristic solutions is generally lower than that of exact solutions. Given the tradeoff that exists between economic cost and power consumption in the TFAP these values are expected. Thus, the loss in economic cost is to a certain extent compensated by the improvement in power consumption. Columns five and six present the average execution time of the heuristic algorithm and the *ILP solver* for solving the various problem instances on a standard workstation with a 2.2GHz dual core processor and 2GB of RAM. The algorithm proposed in this thesis has an order of magnitude lower execution time that ensures fast response and allows its deployment to mobile devices with limited processing power capabilities.

Figure 3.2 presents the cumulative distribution of percent approximation error values for the entire set of problem instances generated by the various experiments. The graph shows that 77% of problem instances are solved with a percent approximation error in economic cost lower than 10%, while 84% of them (65% of total) have percent approximation error lower than 5%. On the other hand, the probability of having a percent approximation error higher than 40% is 2.5%.

A detailed presentation of the distribution of approximation error values e_m per experiment is included in Figure 3.3. Specifically, the diagram includes a different curve for each value $\{1, ..9\}$ of the Experiment Id axis (x axis), that correspond to the distribution of approximation error values e_m of the experiments included in Table 3.5. The projections of the distribution curves in the xy plane present the approximation error values e_m that correspond to the points 40%, 60%, 80% and 95% of the cumulative distribution. The graph shows that 80% of problem instances with low to medium size are solved with relatively

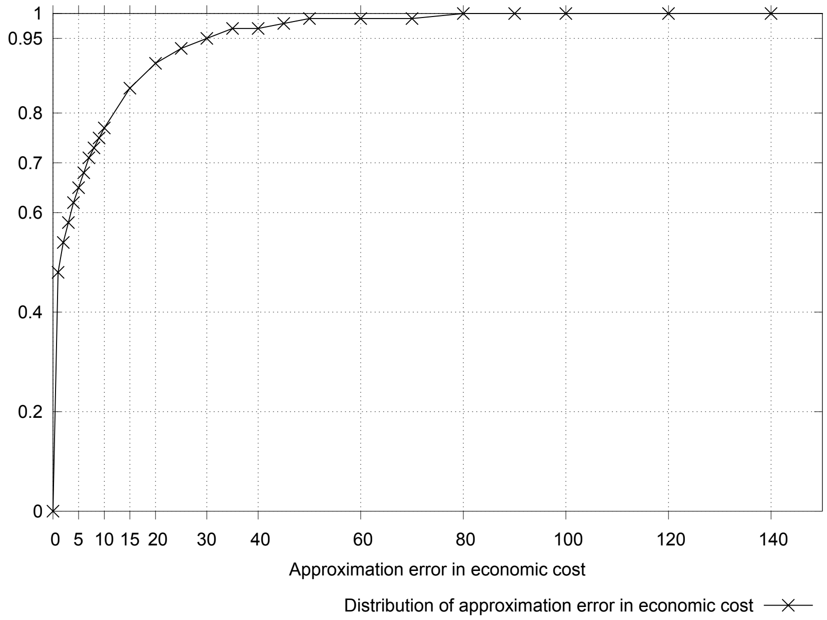

Figure 3.2: Distribution of percent approx. error in economic cost over all problem instances.

high accuracy (15%), while about 75% of them (curve at 60% of cumulative distribution) are solved with worst case approximation error lower or equal to 6%.

With regard to the problems that are generated on the basis of use case scenario UC_3, their exact solution can be obtained by solving relatively large ILP problems [2], with ILP solver execution times that range from several seconds to hours. As the process of generating a sufficient number of problem solutions was rather time consuming, the solutions of a relaxation of the TFAP were used as benchmark. This relaxation is obtained by removing the flow integrality constraints and, thus, allowing the distribution of each flow to more than one interfaces. The resulting problem is a Mixed Integer Programming problem that is generally easier to solve than the pure Integer Linear Programming (ILP). A mixed integer solution of a problem instance is better or equal to the respective integer solution. Thus, the approximation error e_m of heuristic solutions for the experiments 10,11 and 12, as presented in table 3.5 and Figure 3.4, is overestimated. Nevertheless, given the problem size, the results are rather satisfactory as 90% of problem instances for each experiment have approximation error less than 29%.

---

[2] Based on the problem formulation described in [ZG08] each problem instance comprises 4752 variables and 359 constraints.

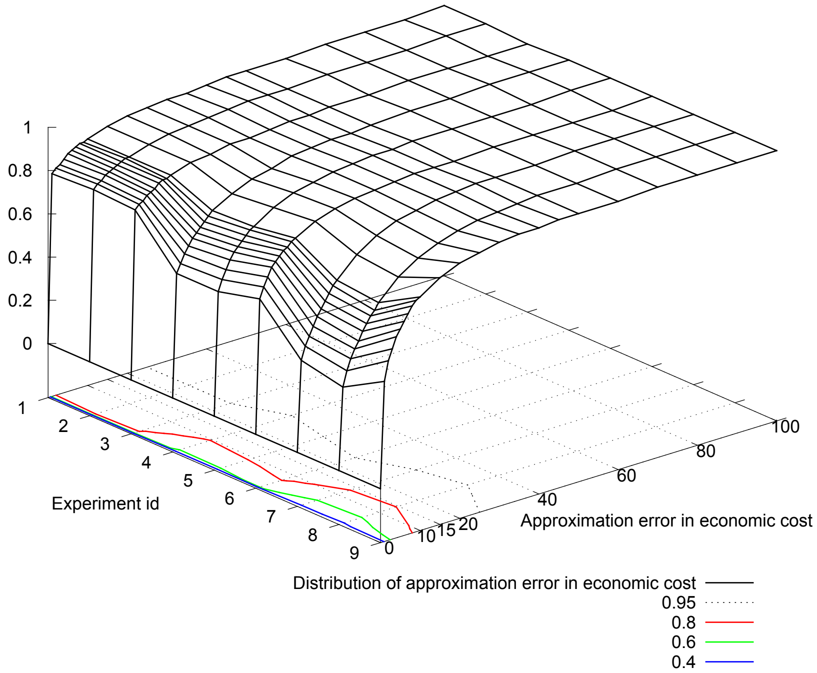

Figure 3.3: Distribution of percent approx. error in economic cost for experiments 1-9.

The evaluation of the impact of (a) random restarts and (b) alternative objective functions, to the approximation performance of the proposed heuristic algorithm, has been performed through solving the problem instances of experiments 1-12 with different algorithm variations. The base version of the proposed heuristic algorithm will be henceforth referred to as LS/RS (Local Search with Restarts) algorithm. The problem instances of experiments 1-12 are also solved with (a) a limited version of the base algorithm without restarts, LS algorithm, and (b) a version of LS/RS that employs the “greedy” objective functions f', g' that were introduced in Section 3.3.1.

Figure 3.5 presents the average percent approximation error in economic cost over all problem instances for the three algorithm variations. The LS/RS algorithm outperforms the other two algorithms, while the worst performance is exhibited by LS/RS with “greedy” objective functions. The increased approximation error of LS algorithm is due to its entrapment to local minima and the absence of any mechanism for recovery from them. On the other hand, LS/RS with “greedy” objective functions algorithm’s poor performance

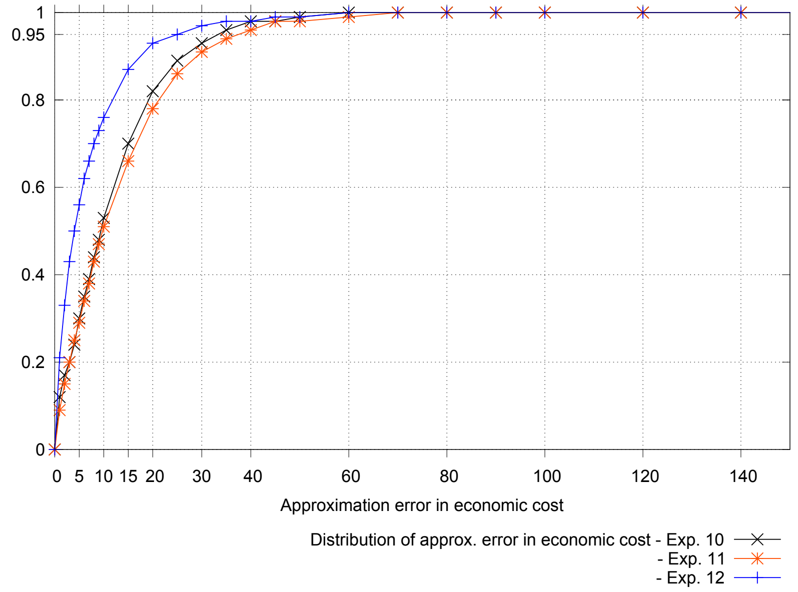

Figure 3.4: Distribution of percent approx. error in economic cost for experiments 10-12.

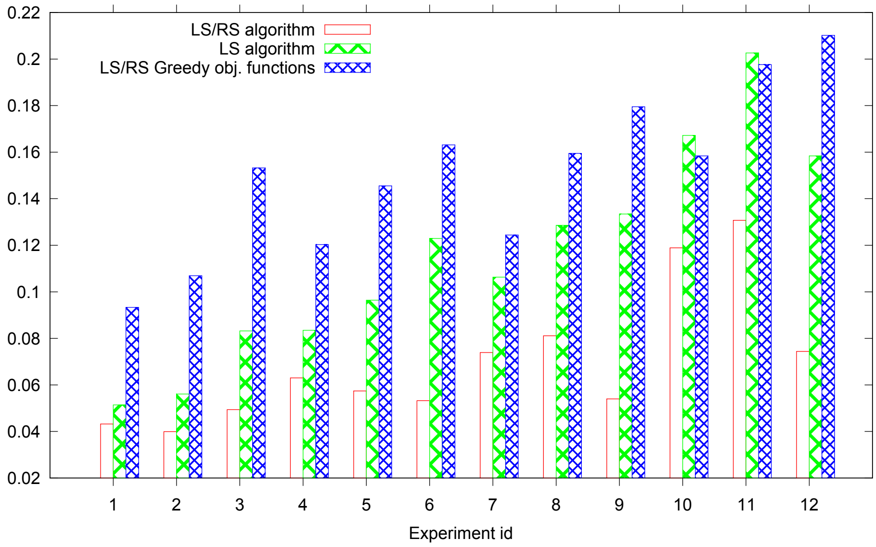

Figure 3.5: Average approximation error in economic cost for 3 algorithm variations.

is caused by the frequent interchange of the economic cost and power consumption minimization search iterations (see Algorithm 2). Specifically, the "greedy" minimization of the economic cost objective, without taking into account power consumption, leads to fast

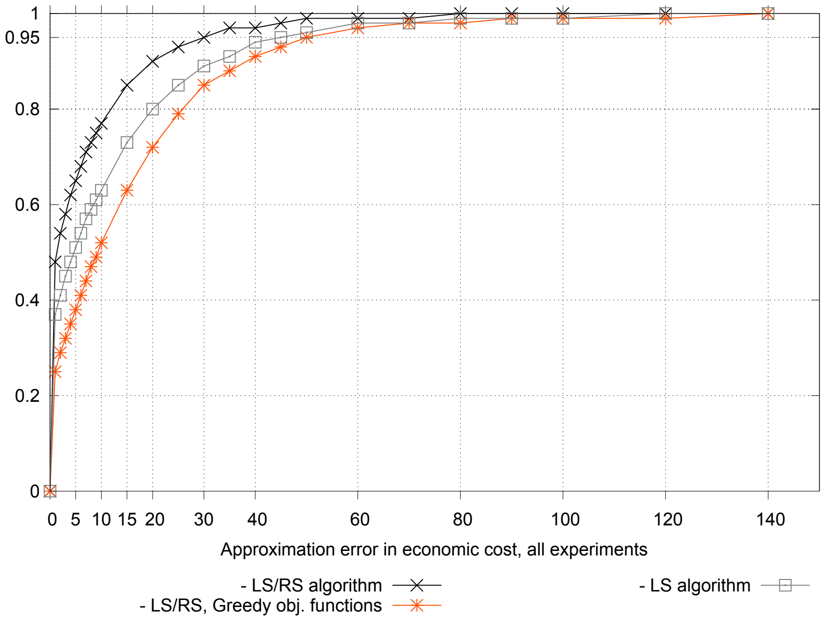

Figure 3.6: Distribution of percent approx. error in economic cost, 3 algorithm variations.

crossing of the power consumption limit $P_{\max}$ after a limited search of the problem state space before $P_{\max}$. On the other hand, the "greedy" minimization of power consumption, results to high economic cost states after reaching $P_{\max}$.

The distribution of the approximation error in economic cost, for the three algorithm variations, is depicted in Figure 3.6. The figure presents results from all problem instances generated by experiments 1-12. A more detailed view on the algorithms' performance is included in Figures 3.7, 3.8, where the distribution of the approximation error concerns the sets of experiments 1-9 and 10-12 respectively. It is clear that the decline of LS and LS/RS with "greedy" objective functions algorithms' performance over LS/RS expands with the increase on the size of problem instances.

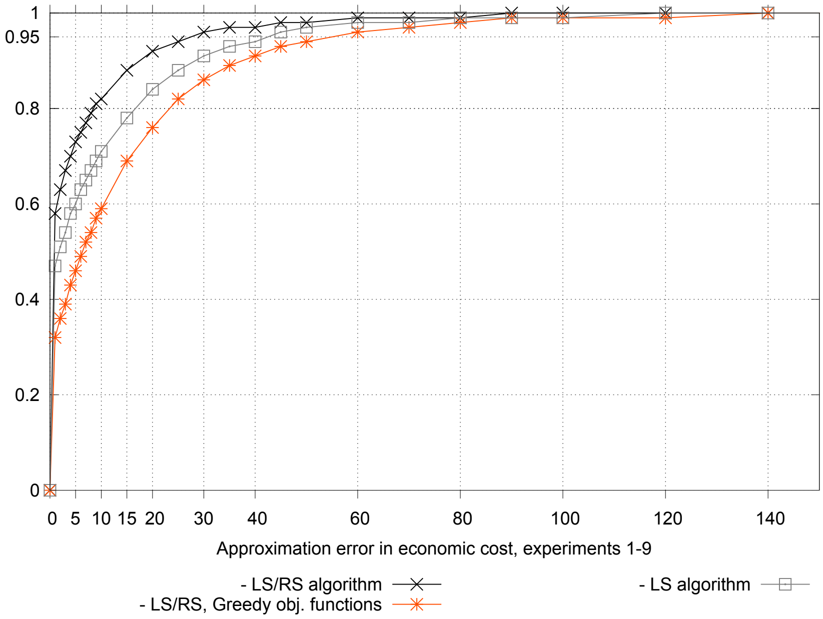

Figure 3.7: Distribution of percent approx. error in economic cost, 3 algorithm variations, experiments 1-9.

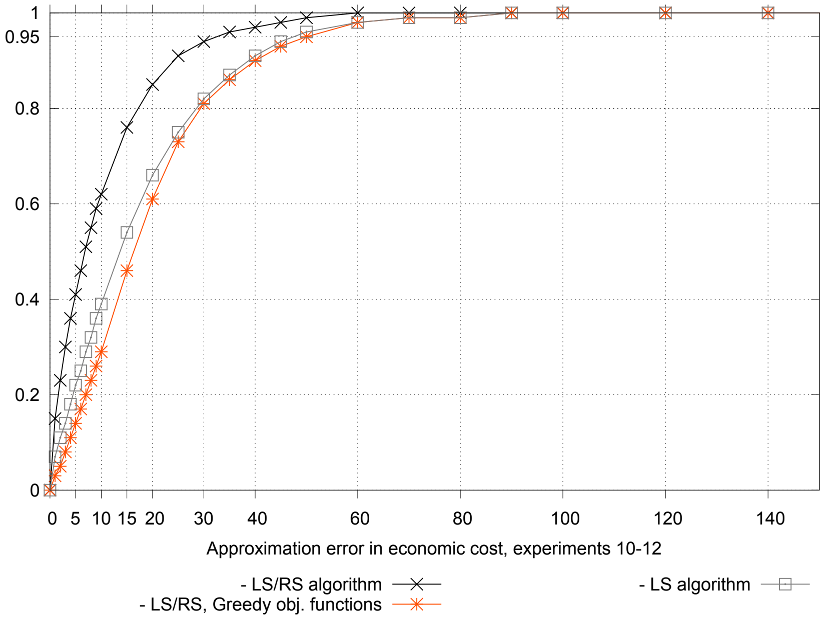

Figure 3.8: Distribution of percent approx. error in economic cost, 3 algorithm variations, experiments 10-12.

3.4.2 Performance evaluation in a dynamic simulation environment

The scope of this work lies on the specification and evaluation of a decision mechanism for optimized utilization of connectivity and energy resources of a multi-homed MMT. The MMT serves application traffic flows of one or more users and operates in a dynamic environment. Thus, the decision context is not static but instead depends on changing application requirements and available radio access resources. In this section an evaluation through simulation is presented with focus on the merits and performance implications of the decision mechanism when applied to the MMT over a specific time horizon. Issues related to the actual enforcement of flow assignment decisions, that involve execution of handovers and flow redirections, are complementary to this work and out of scope of this work. For this reason, as well as for reasons of simplicity and execution efficiency, the evaluation is based on a discrete event simulation system that is driven by events at flow, rather than packet granularity. The focus of the evaluation is threefold: (a) estimation of the approximation error of the proposed heuristic algorithm ALG_h when taking into account the domain of time, (b) comparison of ALG_h with an extended version of an algorithm ALG_u that employs utility functions for network selection in a multi-mode device [NVAGD08], and (c) estimation of the mobility management overhead associated with each algorithm.

The discrete event simulator is implemented in Java and models the random arrival of events that correspond to: (a) arrival and termination of application traffic flows and (b) changes in the available capacity of a set of radio access networks. The traffic flows are generated by a number of users, that engage in video-conference, FTP upload/download and HTTP sessions. Session arrival for each user follows a Poisson arrival process with a rate that depends on the application type. The duration of video-conference sessions is exponentially distributed with mean value 5 minutes and their arrival rate is 2 sessions/hour. Each non real-time (NRT) session comprises a number of packet calls that is geometrically distributed with mean 5. The arrival rate for NRT sessions (FTP, HTTP) is 6 sessions/hour for each session type. The data volume and inter-arrival times of packet calls within a NRT session are randomly generated according to [3GP04].

Each video-conference session corresponds to a pair of (incoming/outgoing) audio flows, each one requiring 12kbps of bandwidth capacity, and a pair of 128kbps video flows. Regarding NRT sessions, each packet call p_z with data volume v_z is mapped to a traffic flow f_z. The bandwidth b_z of f_z is set to the minimum required bandwidth for meeting a maximum tolerated file download time d_z. Thus, $b_z = v_z/d_z$, where d_z depends on the application type and the data volume v_z. In the simulations, $d_z = 3s$ is set for HTTP flows, while for FTP flows d_z is set to: (a) $45s, v_z \leq 2.5MB$, (b) $120s, 2.5MB < v_z \leq 5MB$ and (c) $240s, v_z > 5MB$. Note that in a real setting d_z will probably be part of user preferences.

The time variation of each RAN's available capacity is modeled according to the methodology used in [SNW06] for the evaluation of network selection algorithms. Specifically, each RAN is modeled as a Markov chain with 7 states and each state corresponds to a level of capacity availability, characterized by available bandwidth and access delay. State transitions occur according to the state transition matrix used in [SNW06]. The intervals between state transitions are randomly generated and follow an exponential distribution. Each state transition corresponds to a simulation event that, along with flow arrival and flow termination events, trigger flow assignment decisions.

The simulation system models the operation of a multi-homed mobile router, deployed in a vehicle (e.g., a sightseeing bus) that moves with relatively low speed in an urban area. The mobile router is equipped with 2 UMTS and 2 WLAN (IEEE802.11a/b) radio interfaces. It is assumed that throughout the simulation duration the mobile router's location is constantly served by 4 UMTS, 5 IEEE 802.11a and 5 IEEE 802.11b access networks. The charging rates of UMTS networks are higher than those of the IEEE 802.11a/b networks. Specifically, the rate for each UMTS network is fixed throughout the simulation duration and ranges from 6 to 9 monetary units per kbit, while for WLANs the charging rates range from 1 to 5 units per kbit. The router's served traffic comprises traffic flows that arrive randomly from 5 users that engage in web browsing, FTP and video-conference sessions.

Each simulation execution corresponds to 3 hours of simulated time and focuses on the minimization of the totally incurred economic cost. The randomly arriving flow and network events are processed by the ILP Solver, ALG_h and ALG_u. Each algorithm maintains its own, probably different, flow assignment state that is updated after each execution. The ILP Solver decisions result to the minimum possible cost for a given event sequence. The proposed heuristic algorithm ALG_h is also compared with a network selection algorithm ALG_u that takes into account multiple criteria of different priority, evaluated by utility functions [NVAGD08]. The criteria considered in this evaluation are: (a) cost C_c, (b) power consumption gain C_{ge} [NVAGD08], (c) available bandwidth C_b, (d) required QoS class C_q. The priorities of C_c, C_{ge} are set to 3 (high), 0 (ignored) respectively for all simulation executions. Other priority levels used in [NVAGD08] are 2 (medium) and 1 (low). The priorities of C_b, C_q depend on the served traffic and their configuration is explained below.

Note that ALG_u, as specified in [NVAGD08], focuses on the selection of a single RAN, while for capacity or economic efficiency reasons more than one radio interfaces may need to be activated. In this evaluation ALG_u is used for the selection of more than one RANs by applying it iteratively for the assignment of flows of QoS class C (conversational) and then for S, I and B classes respectively. The iteration for each QoS class Q_i involves the following simple steps:

1. Set the priority of C_q to 3, 2, 1 or 0 according to the current value of Q_i (C, S, I or B respectively).
2. Set the priority of C_b on the basis of the ratio of total required bandwidth of Q_i class flows to the available capacity through all radio interfaces for traffic of class Q_i. This ratio is normalized and mapped to the allowable priority values $\{0, ..., 3\}$.
3. Evaluate available bearers with QoS class better or equal to Q_i according to ALG_u.
4. Assign flows of class Q_i to the first possible bearer(s).

Figure 3.9 presents the percent approximation error in economic cost and energy consumption of ALG_h and ALG_u, as it is averaged on 100 simulations. The performance of each algorithm is compared against the economic cost and energy consumption incurred by executing the ILP solver and enforcing its decisions upon all simulation events. Regarding the execution frequency of ALG_h and ALG_u, a different policy has been adopted that is more appropriate for a real-world deployment, where optimality needs to be balanced with system stability. Specifically, each execution of ALG_h or ALG_u is followed by a phase where new events are handled with the least possible modifications to the current state. For instance, on a flow arrival event the new flow is assigned to the cheapest radio interface with spare capacity, on a flow termination event no action is taken etc. This phase ends with the arrival of any event that invalidates the current flow assignment, e.g., the QoS of a used bearer service violates its flows' requirements or assignment of a new flow is not possible. A new valid flow assignment is derived through algorithm execution and the process continues. As illustrated in Figure 3.9, ALG_h provides good approximation to the optimal cost, despite the aforementioned conservative execution policy. Due to the large decline from optimal economic cost, ALG_u outperforms ALG_h in terms of power consumption.

The average mobility management overhead caused by the application of the algorithms' (ALG_h, ALG_u) decisions in each simulated scenario is depicted in Figure 3.10. The figure

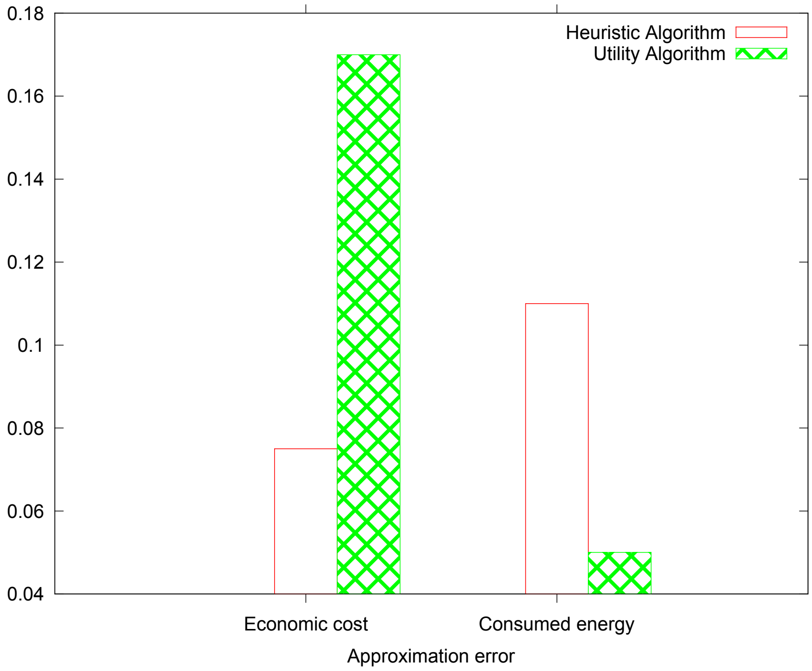

Figure 3.9: Average approximation error in economic cost and consumed energy.

presents the average number of flow redirections and horizontal handovers per minute triggered by each algorithm. A flow redirection refers to the change of the serving radio interface of an active flow, while a horizontal handover occurs when an active radio interface changes its point of attachment to a different RAN. The graph shows that the overhead of ALG_h is very close to the respective overhead of algorithm ALG_u, that represents a simpler and more straightforward flow assignment scheme. In any case, the incurred mobility management actions are limited, given the gain in economic cost and the number of concurrently served users.

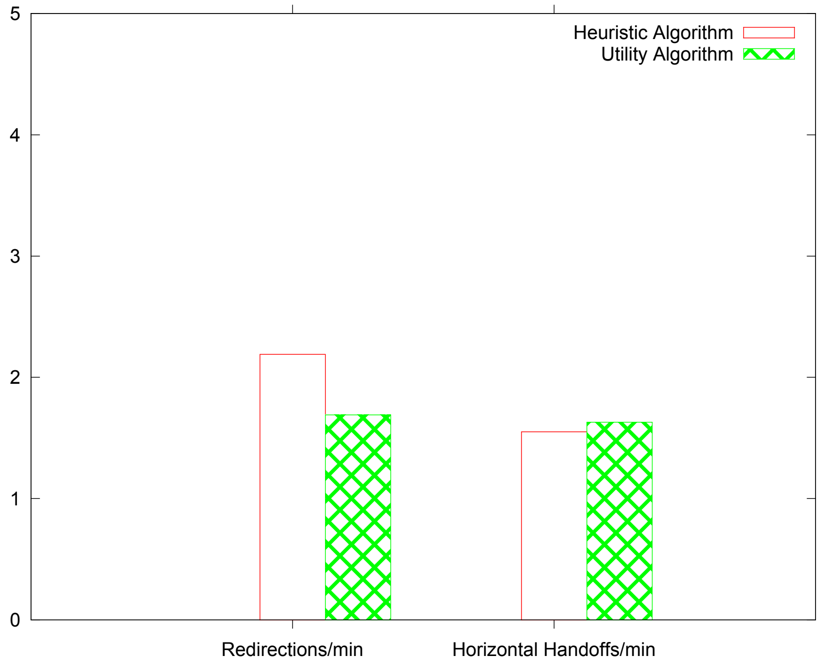

Figure 3.10: Average number of flow redirections and horizontal handovers per minute.

Chapter 4

An Agent-based Architecture for Handover Decision Support in Heterogeneous Networks

4.1 Introduction

The fourth generation (or Beyond 3G-B3G) of mobile communications systems is orientated towards the integration of available wireless access technologies. The envisaged architecture of a 4G network comprises a variety of wireless access (WLAN, Bluetooth etc.) and cellular 3G or 4G networks interfacing with access routers to a common core IP network, that serves their interconnection and integration with earlier generation networks (PSTN, ISDN, 2G/2.5G etc.) through appropriate gateway routers [Zah03]. Handover management extends its scope in 4G targeting seamless mobility across cells: (a) of the same or different technology (horizontal vs. vertical handover), (b) operated by the same or different providers (intra-domain vs. inter-domain handover). Thus, additional factors need to be taken into account during handover initiation and decision, such as user's cost tolerance and contractual constraints, applications' requirements and priority, terminal device status (battery status, available radio interfaces etc.) [SZ06]. In this context the role of handover is being enhanced from maintaining connectivity to optimizing connectivity.

The sources of handover triggering events, as well as of the information required for handover decision, do not lie exclusively on the network or link layers. Instead, given that application and user related factors will be taken into account, higher layers, such as the application layer, would also have substantial contribution. Thus, an application layer approach is required for a handover initiation and decision solution, as it is also proposed in [GJ03] where the requirements for an Always Best Connected (ABC) service are set.

From a user viewpoint, a technical solution for handover initiation and decision enables efficient connectivity in a context of multiple wireless access networks. However, from a network operator's perspective such a solution acts as a market enabler that allows clients to either join or leave its network depending on their current utility. A trustworthy implementation of this capability cannot be offered by a single network provider. The reason is that it has no incentives to provide reliable and in-time information regarding available wireless networks and consequently let its customers utilize third-party services. A viable solution should: (a) incorporate various wireless operators, (b) support market competition through easy integration of new entrants, (c) adopt a common, unambiguous information schema for interoperability of the exchanged information (e.g., descriptions of network capabilities, so as to enable effective decision making), (d) build on a commonly accepted model of trust relationships so as to be relied upon by users and network operators.

In this section an agent-based architecture is proposed for handover management in 4G with main focus on the initiation and decision phases of the handover mechanism. The architecture is specified in terms of (a) definition of agent types and their deployment in user terminals and access networks, and (b) description of agent collaboration for determining the timing and target of handovers. Moreover, a study of the architecture's integration into a heterogeneous network setting, comprising 3G cellular networks and IEEE 802.11x WLANs, is provided. Finally, performance implications of the proposed architecture are analyzed on the basis of results drawn from a simulation system. The rest of this section is organized as follows: Section 4.2 analyzes the business environment where the architecture will be integrated. An overview of the architecture along with the merits of an agent-based approach in handover management is presented in Section 4.3. Section 4.4 focuses on agent collaboration during handover management. A performance evaluation of the proposed architecture is presented in Section 4.5.

4.2 Business actors and their relationships

The architecture focuses on inter-domain handovers, as well as on vertical intra-domain handovers. The assumption is that horizontal handovers in the bounds of a domain are transparently handled by the network's mobility management mechanism. In the case of a single network provider, available Radio Access Networks (RANs) may be interconnected through a common core network. However, in this thesis the focus is on the more general case where RANs provide access to different core networks, interconnected either directly or through the public Internet. Two types of network providers are assumed: (a) core network operators that manage zero or more possibly heterogeneous RANs attached to their core network, (b) RAN operators that manage one or more RANs attached to the networks of core operators.

End-users gain access to the services of various wireless providers through a single subscription. This is enabled by a business entity that maintains roaming agreements with the network operators and offers a value-added service on top of their network infrastructure. Its role is analogous to that of an ABC service provider, as defined in [GJ03]. However, it will be referred to as Multi-Access Provider (MAP) as the focus is on a subset of the ABC service capabilities. MAP subscribers are not required to maintain contracts with network providers-partners of the MAP. The MAP is responsible for their authentication and bills them by aggregating their respective charges from the various network providers. In addition, it supports handover management by providing a platform where (a) network operators publish their services and (b) user terminals discover network services and select the most appropriate on the basis of user, device and application related constraints.

However, MAP cannot guarantee the accuracy of information published through its infrastructure. For instance, a network provider may publish service descriptions that do not correspond to their actual, probably inferior, performance. In order to discourage providers from abusing the infrastructure, the presence of the regulatory authority is deemed important. The regulator is trusted by all actors and enhances the credibility of offered services. Its role will be further analyzed in Section 4.3.3.

4.3 Architecture Overview

4.3.1 An agent based approach

The use of software agents facilitates the deployment, maintenance and management of the system. The capability of a network node to host the execution of system components depends on the presence in it of an agent platform. Once platforms are deployed, agents can populate them either by self-motivated mobility or remote instantiation with the use of appropriate management software. In the same way new software versions can be deployed. Management is enabled by management capabilities provided by the platform through a standard interface (FIPA compatible agent platforms) [FIP10].

Interoperability of system components is a critical design issue as they will be implemented by different parties (MAP, regulator, network providers, other application service providers) with possibly different perspectives on the problem domain. A straightforward approach for interoperability is based on the definition of Application Programming Interfaces (APIs) and messaging protocols. Such specifications are implementation specific, focus on the syntax of the messages and may have different interpretation by the various providers. Agent interoperability is based on the exchange of messages expressed in one of the widely accepted Agent Communication Languages (ACLs), FIPA ACL [FIP02a] or KQML [LFP99]. An ACL message encapsulates the communication payload and describes it in a domain independent way with a predefined set of attributes. The payload is expressed in a content language (e.g., KIF, RDF) with the use of vocabulary from shared ontologies. Consensus, thus, among providers is reached at a higher, conceptual level that ensures unambiguous interaction.

4.3.2 MAP Network Architecture

The specification of the proposed architecture, as well as a study of its integration in the infrastructure of current wireless networks, will be presented with reference to a 4G heterogeneous network setting. For the sake of simplicity three types of wireless providers will be considered: (a) 3GPP (GSM/UMTS) operators, offering packet switched as well as circuit switched services, (b) WLAN providers, operating IEEE 802.11x WLANs for Internet access in public hot spots and (c) Interworking-WLAN (I-WLAN) providers.

An I-WLAN provider manages a WLAN RAN that interworks with the core networks of one or more 3GPP systems, henceforth referred to as 3GPP Public Land Mobile Networks (PLMNs) or simply PLMNs. Recall that the 3GPP-WLAN interworking architecture has been specified by 3GPP [3GP08a] with purpose of (i) enabling WLAN public access to subscribers of PLMN operators and (ii) enabling WLAN access to IP-based services of the 3GPP PS domain. An I-WLAN allows a mobile terminal to be authenticated by its home PLMN (HPLMN) and use packet switched services of the HPLMN, other visited PLMNs (VPLMNs) or directly connect to the public Internet via the I-WLAN network depending on user subscription terms.

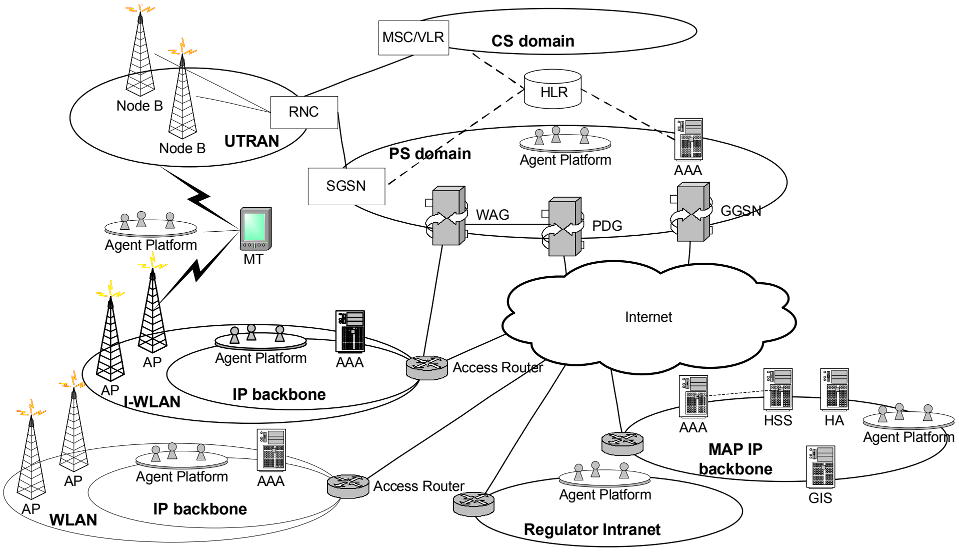

Figure 4.1: MAP network architecture

Figure 4.1 depicts a simple 4G setting that will be used as a case study for the description of various operational scenarios of the proposed approach. In this setting, the MAP cooperates with a WLAN, a UMTS (UMTS$_1$) and an I-WLAN operator. The latter provides access to core services of UMTS$_1$, in addition to basic Internet access. Moreover, it is assumed that the coverage areas of all wireless access networks intersect in the user's geographical region. A contract with the MAP enables a user's access to services offered by all MAP partners. Each wireless provider and the regulator incorporate an agent-platform in their core networks in order to participate in the architecture. Especially for operators of networks with broad geographical coverage, the management of more than one instances of an agent-platform may be required for efficiency and load distribution purposes. Each instance can serve the users of a geographical region, e.g., a city or a more restricted area.

The MAP's network infrastructure is also outlined in Figure 4.1. It comprises an IP backbone that interconnects a set of core elements. These core elements, accessible by terminals and wireless networks through the public Internet, include:

- a Geographical Information System (GIS) for location-based retrieval of information related to the type and coverage of its partner- networks,

- a Mobile IP Home Agent (HA) for macromobility support,

- an Agent Platform hosting agents that wrap services offered by other core elements and make them accessible to agents distributed in user terminals and access networks,
- a Home Subscriber Subsystem (HSS) accessible by HSSs of 3GPP networks for authentication and authorization purposes, call routing and user profile retrieval, and
- an AAA server that interfaces with respective AAA servers of network providers using RADIUS or Diameter. It is referenced by them for user authentication and authorization, while it concentrates their charging records for billing purposes. The AAA server interfaces and acts as a proxy to HSS that is the primary source of authentication, profile and authorization information.

Concerning user terminal equipment, a MMT is assumed, capable of connecting to both IEEE 802.11 and UMTS networks. A basic requirement for a MMT is the integration of a UICC smart card with USIM/SIM modules for authentication with UMTS/GSM networks respectively. UICC smart cards are owned by the MAP and include each subscriber's International Mobile Subscriber Identity (IMSI) that is also issued by the MAP.

4.3.3 Agent types and functionality

Figure 4.2 depicts the architecture's agent types, along with the agent platforms that host their execution. Agents' different shading is indicative of the authority they represent. Platforms comply with the FIPA agent management reference model [FIP10] that specifies three basic logical components: (a) Agent Management System (AMS) agent, (b) Directory Facilitator (DF) agent and (c) the Message Transport Service (MTS), the default communication method between agents on different platforms. AMS is a mandatory component that provides agent registration, life cycle control and white pages services, while DF is an optional component that provides yellow pages services to the agents of a platform. The ownership of AMS and DF correlates with the administration of the platform. Agent communication is based on ACL messages expressed in either FIPA ACL or KQML [LFP99]. Message transport is carried out over TCP/IP with use of HTTP, IIOP, WAP etc.

Multi-access provider agents

They wrap services offered by the MAP's core elements and provide them via an ACL interface to other agent types executing in wireless networks and user terminals. Among the main support services offered by the MAP agents are:

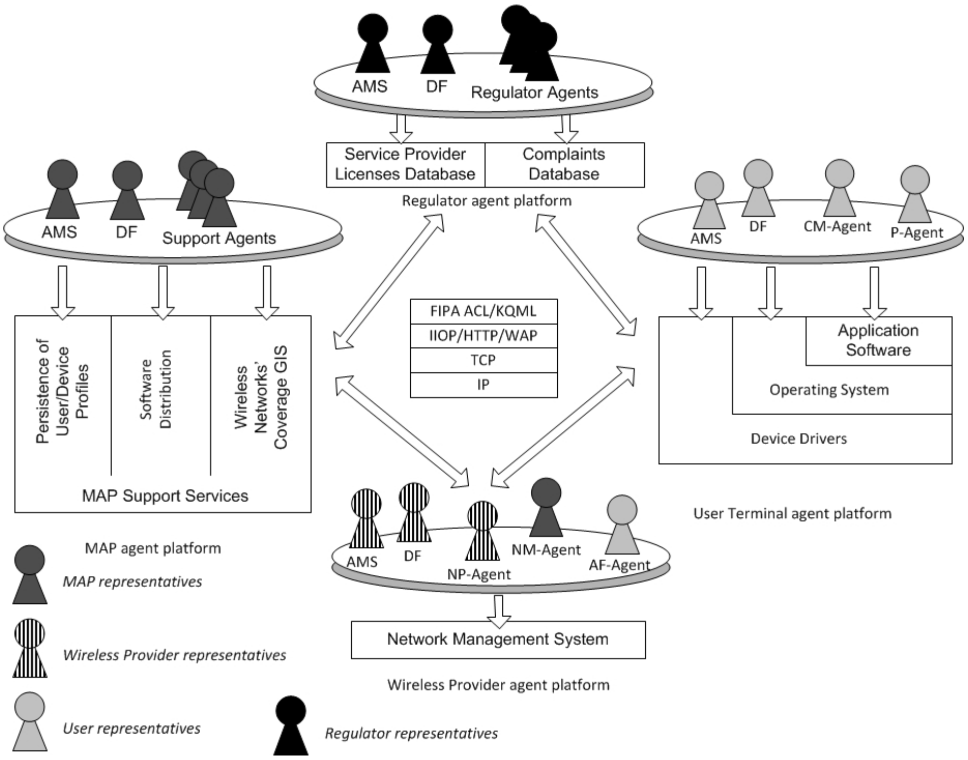

Figure 4.2: Agents comprising the architecture.

- software distribution, where latest versions of drivers and software libraries are distributed on demand in order to enhance the utilization of the terminals' network interface(s),

- management of user profile information, that is transferred to authorized agents in order to reason and act upon it, and

- location-based retrieval of network information. The geographical coverage of each RAN as well as information concerning its type and offered services are stored in MAP's GIS. MAP agents interface with the GIS and serve requests on (a) the wireless networks of certain type that cover a geographical location and (b) the wireless networks that have overlapping coverage and the same type with a given network.

Wireless provider agents

A wireless provider's platform executes agents that represent all business actors except for the regulator. The platform includes one Network Provider agent (NP-agent), one Network Monitor agent (NM-agent) and several Access Facilitator agents (AF-agents) representing the wireless provider, the MAP and currently connected users respectively. *NP-agent* provides an ACL interface for controlled access to network management information of the current network. The information exposed by the NP-agent includes descriptions of information transfer services provided by its network, characterized by various attributes such as QoS, cost etc. *NM-agent* aggregates information regarding access networks that intersect with the current network's coverage area. This information is retrieved through subscription to respective NP-agents. Each NM-agent serves numerous *AF-agents*, corresponding to the current network's users. AF-agent is a user proxy responsible for handover initiation and decision and subscribes to NM-agent for receiving information on networks with presence to the MMT's current location. It incorporates user and terminal profile information relevant for its decision making. In order to minimize network latency in its communication with the terminal's agents, AF-agent migrates and executes each time in the platform that corresponds to the current access network.

Terminal device agents

Each terminal device utilizing the MAP services, hosts two agent types that are user representatives, Profile agent (P-agent) and Connection Manager agent (CM-agent). *P-agent*'s role is to communicate the perceived QoS, user preferences and application requirements to AF-agent in order to make informed decisions on the target and timing of handovers. *CM-agent* is responsible for successful execution of handovers, initiated by the user's corresponding AF-agent. Handover execution is preceded by a procedure that determines the terminal's capability (in terms of software requirements, RSS) of accessing the selected network. Note that the terminal's hardware configuration is taken into account during handover decision by AF-agent that informs CM-agent of the appropriate driver versions and protocol implementations that the terminal should support in order to connect effectively to the specified network. CM-agent checks the terminal's software configuration and reports any deficiencies to a MAP agent in order to download updated versions.

Regulator agents

The software agents that execute on the regulator's platform wrap appropriate databases, and either update or make their contents available to other agents. Two types of databases are considered: (a) a database DB_l with service licenses granted to the various network operators, (b) a customer complaints database DB_c. On the basis of DB_l contents, regulator agents provide information regarding network providers and the services they are licensed to offer. Such information is requested by NM-agents whenever a new type of service is retrieved from a NP-agent. Moreover regulator agents, update DB_c on the basis of notifications related to misleading service descriptions published by providers. Such notifications are sent by P-agents whenever the QoS of a service declines significantly from that advertised. These notifications are analyzed by the regulator and, if necessary, further investigation is conducted. DB_c is also updated with notifications regarding locations with high radio interference. P-agents send such information whenever high interference is perceived by the MMT. The regulator can check areas with high concentration of notifications for operation of unlicensed antennae.

4.4 Handover Mechanism

4.4.1 Network Provider Platform Bootstrap

Bootstrapping of a network provider's platform involves authentication of the platform's NP-agent with the MAP and download of the latest version of the NM-agent software. Figure 4.3 presents the message exchange during I-WLAN's agent platform initialization. After its instantiation (2), NM-agent requests from the local NP-agent geographical information regarding the coverage area that this platform serves (3). On the basis of such information NM-agent retrieves from the MAP the addresses of NP-agent$_{WLAN}$ and NP-agent$_{UMTS-1}$ that correspond to networks that intersect with the given coverage area (4). Finally, NM-agent retrieves bearer service descriptions from all relevant NP-agents through subscription (5, 6, 7). Adoption of a subscription model allows only modified service descriptions to be propagated to NM-agents, thus minimizing network traffic that could be generated by periodically polling NP-agents for their offered services.

4.4.2 Handover Initiation

Handover initiation is an iterative task that triggers handovers whenever the terminal's current connection(s) do not meet the applications' requirements in terms of bandwidth, QoS, security etc. P-agent perceives and makes available to AF-agent a series of handover initiation events such as:

- link quality degradation on an active radio interface,
- change to connectivity requirements due to the start or termination of an application,
- change to the status or configuration of the terminal device, e.g., low battery level, availability or unavailability of a radio interface,
- special core network capabilities required by an application, for instance sending a MMS message requires access to the core of a 2.5/3G network,
- discovery of a new access network.

AF-agent also perceives handover initiation events from NM-agent whenever the status of a subscribed network (e.g., congestion level) changes.

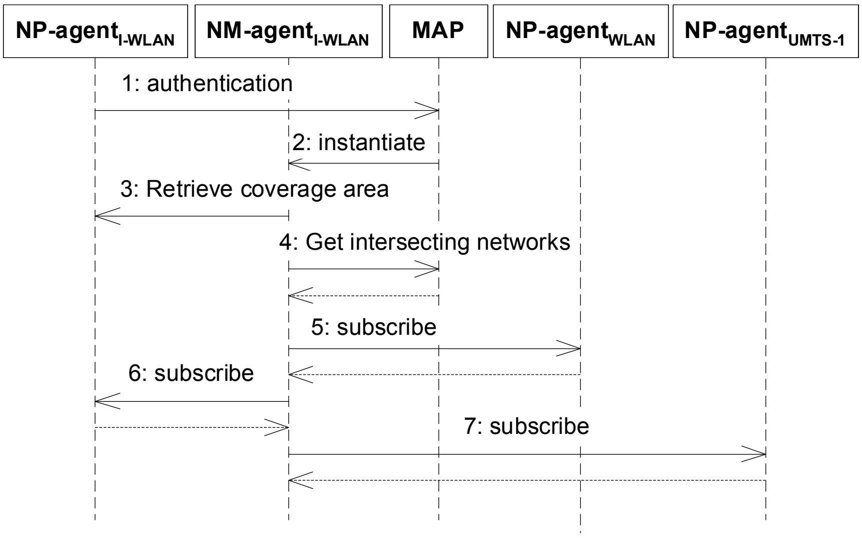

Figure 4.3: Bootstrapping of network provider's platform

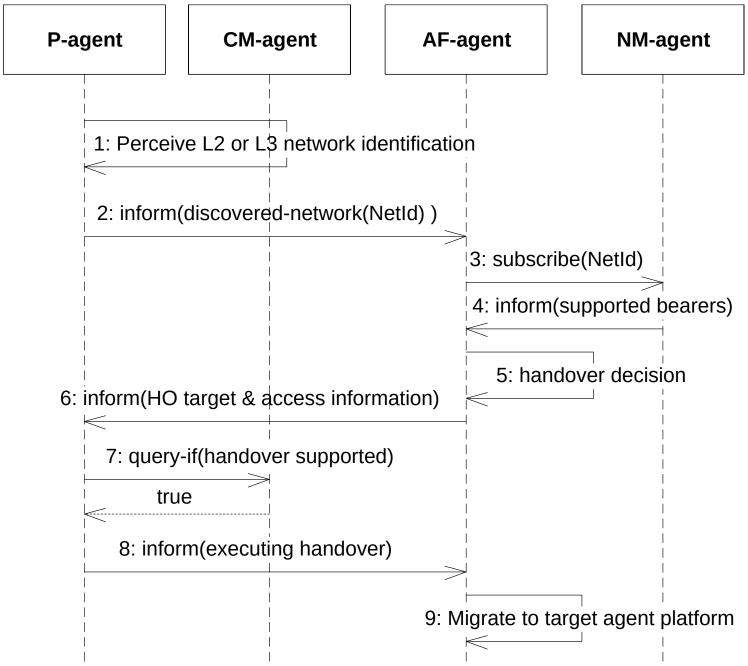
P-agent
CM-agent
AF-agent
NM-agent
1: Perceive L2 or L3 network identification
2: inform(discovered-network(NetId))
3: subscribe(NetId)
4: inform(supported bearers)
5: handover decision
6: inform(HO target & access information)
7: query-if(handover supported)
true
8: inform(executing handover)
9: Migrate to target agent platform

Figure 4.4: Handover initiation and decision

Agent interaction in a scenario of new wireless network discovery is presented in Figure 4.4. It is assumed that the MMT is initially connected to the UMTS network. As it approaches a hot spot covered by the WLAN provider, its WLAN interface starts receiving IEEE 802.11 beacon frames from WLAN APs. Each beacon frame incorporates a SSID information element that identifies the WLAN provider. This information is extracted and made available to P-agent through an appropriate event (1). However, besides network identification, beacon frames do not include other information attributes that may assist the MMT's handover decision. Such information could be retrieved after associating with the AP, a relatively time consuming task that may prove useless, if finally the WLAN is not selected. In the proposed approach discovery of bearer service information is delegated to the AF-agent. P-agent notifies AF-agent through its current UMTS connection on the perceived event with an initiate handover message that includes the discovered network's SSID (2).

Information on the types of bearer services offered by each wireless network is retrieved by AF-agent through subscription to the local NM-agent (messages 3, 4). On subscription initiation, AF-agent provides to NM-agent an identification of the network that is interested on receiving information about. AF-agent's subscriptions span only wireless networks that are available in the MMT's current location. These networks are discovered by the MMT on the basis of L2/L3 information received by its radio interfaces. As an alternative, MMT's current location, retrieved by a GPS receiver, could be utilized for network discovery. AF-agent *decides* on the most appropriate target for handover (5) and forwards to P-agent all relevant information for handover execution (network type and registration information, protocol versions, authentication type etc.)(6). P-agent requests the execution of the handover from CM-agent (7) that assesses the terminal's capability of connecting to the proposed network. In the positive case, P-agent notifies AF-agent (8) that migrates to the target network's agent platform in order to serve its principal with the minimum network latency (9). Note that in case of multi-homed MMTs, that maintain connections with more than one networks, AF-agent's migration takes place towards the network that provides the lowest communication delay with the terminal's agents.

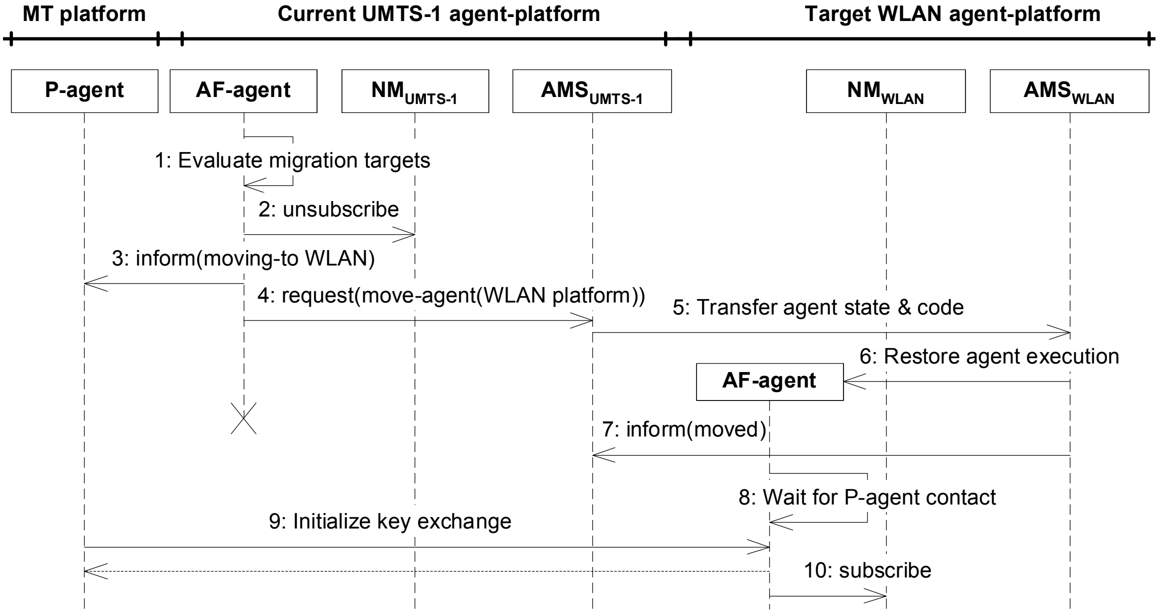
MT platform
Current UMTS-1 agent-platform
Target WLAN agent-platform
P-agent
AF-agent
NMUMTS-1
AMSUMTS-1
NMWLAN
AMSWLAN
1: Evaluate migration targets
2: unsubscribe
3: inform(moving-to WLAN)
4: request(move-agent(WLAN platform))
5: Transfer agent state & code
6: Restore agent execution
AF-agent
7: inform(moved)
8: Wait for P-agent contact
9: Initialize key exchange
10: subscribe

Figure 4.5: AF-agent migration from UMTS-1 to WLAN agent platform.

A more detailed description of AF-agent's migration procedure is depicted in Figure 4.5. Continuing the above scenario, AF-agent decides to migrate to the agent platform of WLAN, that is the MMT's target for handover execution (1). Before migration, AF-agent unsubscribes from NM-agent$_{UMTS-1}$ (2) and informs P-agent of the new platform that will host its execution (3). Agent migration is requested by the current platform's AMS (AMS$_{UMTS-1}$) (4) that transfers agent state and code to AMS$_{WLAN}$ (5). The latter restores AF-agent's execution (6) and informs the former on the migration success (7). After resuming its execution on the new platform, AF-agent waits a contact from P-agent in order to continue its normal operation (8). P-agent communicates with AF-agent once handover execution to the new network is complete. A key exchange takes place among agents in order to prove their identity (9) and finally AF-agent subscribes to NM$_{WLAN}$ (10).

4.4.3 Handover Decision

Recall that the heterogeneous network setting described in Section 4.3.2 assumed a MMT located in an area served by WLAN, UMTS and I-WLAN network operators. In this setting, handover decision extends its scope to the selection of both wireless access and core networks. A basic requirement for optimal selection of access and core networks is a consistent and commonly adopted data model for describing data transfer services. In this thesis, the adopted approach for the description of the various services is based on the categorization and attributes used in [3GP10d, 3GP09f] for the characterization of telecommunication services in a 3GPP PLMN. Specifically, telecommunication services are classified into *bearer services*, that provide information transfer between access points of a network, and *teleservices* that provide a complete capability. The service descriptions published by network providers, through NP-agents, belong to the bearer service category. As handover execution, in most approaches, is handled in the data link (L2) or network layer (L3), handover decision involves selection of L2 or L3 bearer services offered by different networks. Bearer services are described in [3GP10d, 3GP09f] through a set of attributes (see Table 4.1 for a subset of them).

The structure of a bearer service description is depicted in the class diagram of Figure 4.6. The protocol that is used by a bearer service for information transfer as well as a description of its endpoint and its access requirements (e.g., authentication method) are provided by the "Access Interface" service attribute. Its range is set to instances of the Protocol Endpoint class that has the same semantics as the respective class defined in the

Common Information Model (CIM) Network Model specified by Distributed Management Task Force (DMTF) [DMT10]. The Protocol Endpoint (PE) class identifies the address or location where the bearer service is available and incorporates service configuration information (e.g., supported packet size, negotiable QoS attributes, authentication methods). Moreover, the class characterizes the OSI stack layer that the service belongs to. For instance, the LAN Endpoint and IP Protocol Endpoint classes, defined in [DMT10], extend PE and describe L2 and L3 bearer service endpoints respectively.

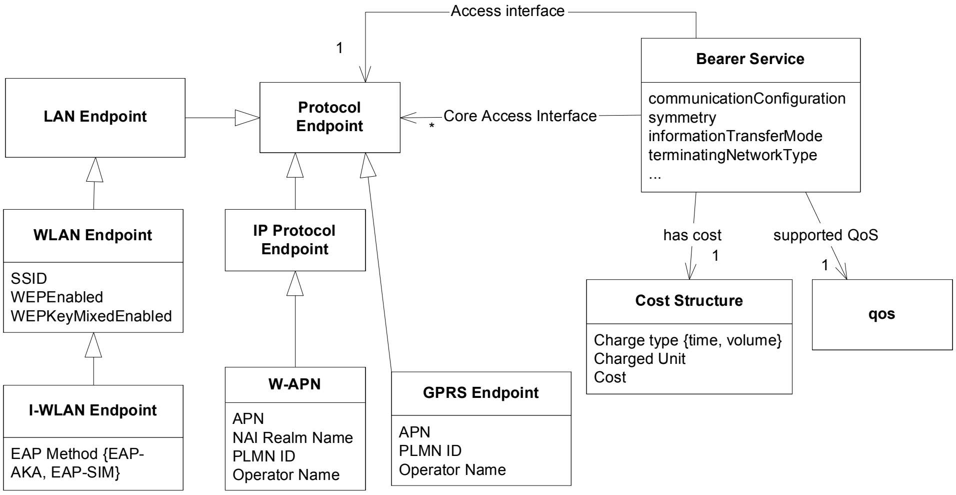

Figure 4.6: Class model of bearer service descriptions.

For the assumed network setting, a set of additional Protocol Endpoint subclasses need to be introduced for the description of available bearer services. These classes, depicted in Figure 4.6, are: (a) I-WLAN Protocol Endpoint that extends the WLAN Protocol Endpoint [DMT10] by appending the "EAP Method" attribute that specifies the protocols that are supported for 3GPP authentication, (b) Wireless Access Point Name (W-APN) that identifies a L3 access point to the PS domain of a 3GPP network interworking with a WLAN (note that instances of W-APN class characterize the "Access Interface" of 3GPP core bearer services) and (c) GPRS Protocol Endpoint that identifies a packet data network in a PLMN.

The types of core networks that are accessible through a given RAN are specified in the "Accessible Core Network Types" attribute, while the endpoints of their respective core bearer services are given in the "Core Access Interface" attribute. The latter is a multi-value attribute and ranges over instances of the Protocol Endpoint class. The bearer service descriptions that correspond to the aforementioned core access interfaces should also be provided by the NP-agent of the wireless network that interworks with them and made available to NM-agents. Finally, the cost of using a certain bearer service is specified in the 'Cost' attribute that ranges over instances of the 'Cost Structure' class.

Table 4.1: Bearer service description

Attribute name		***Values***
Information Transfer attributes	Connection mode	Connection oriented, Connectionless
	Traffic type	Constant or Variable Bit Rate
	QoS	Instance of qos class
	Communication configuration	Point to Point (PTP), Point to Multipoint (PTM), Broadcast
	Symmetry	Unidirectional, Bidirectional Symmetric, Bidirectional Asymmetric
Access attributes	Access interface	Instance of Protocol Endpoint class
Interworking attributes	Type of terminating network	PSTN, ISDN, PSPDN, PDN[1], PLMN, direct internet access
	Accessible core network types	PSTN, ISDN, PSPDN, PDN, PLMN
	Core access interface	Instances of Protocol Endpoint class
General attributes	Cost	Instance of Cost Structure class

On the basis of bearer descriptions retrieved from NM-agent, AF-agent creates the best combination for user access. At first, a selection of Access Bearers (ABs) (i.e., bearer services with an L2 wireless protocol endpoint) takes place that are usable by the MMT's radio interfaces. Next, Core Bearers (CBs) are retrieved that are accessible through the selected access bearers, i.e., bearers that their "Access interface" matches the "Core access interface" of at least one of the selected access bearers. As a result a set of pairs (AB$_i$, CB$_i$) is created. The value of "Type of terminating network" attribute is then checked in order to ensure that the type of termination required by MMT applications is provided by selected bearers. This filtering does not apply to access bearers that interwork with at least one core bearer. The combined cost and QoS of each pair (AB$_i$, CB$_i$) is then calculated and used as input to a TFAP algorithm, as the one described in Section 3.3 or one of the methods proposed in [SJ05] or [XV05].

---

[1] Packet Switched Public Data Network, Public Data Network

4.4.4 Mobile Terminal's Initial Access and Authentication

During initial log on of a disconnected MMT, network selection aided by AF-agent, is not possible. Its instantiation takes place after successful authentication through a network provider supporting the MAP services. In this case network selection is based on a predefined priority list of wide coverage providers. After user authentication, P-agent and CM-agent are instantiated on the MMT's agent platform while an AF-agent instance is created in the MAP agent platform and moves to the platform of the user's current network. The terminal authenticates with the MAP at initial log on and prior to each inter-domain handover. Authentication is carried out by the terminal's Universal Integrated Circuit Card (UICC) that incorporates appropriate software modules such as Universal Subscriber Identity Module (USIM), for authentication with UMTS over UTRAN, and Extensible Authentication Protocol (EAP)-Authentication and Key Agreement (AKA) that allows the execution of the UMTS AKA protocol over a WLAN access network [KH03]. Authentication information is forwarded through the current wireless provider's AAA or HSS to the corresponding elements of the MAP that is responsible for user authentication and authorization.

4.4.5 Handover execution

The proposed approach is transparent to handover execution protocols and various mechanisms, either network, transport or application layer may be employed in order to ensure flow continuity across different RATs or domains. As regarding the incorporation of Mobile IPv6 in the architecture, MIPv6 Home Agent (HA) is situated, as described in Section 4.3.2, in the MAP's domain. Mobile terminals authenticated with the MAP perform binding updates to the HA with their current care-of addresses (CoAs) in order to be reachable by correspondent nodes through their public Home Address (HoA). Routing optimization should be applied where possible (it depends on the capabilities of the correspondent node) so as to avoid transforming the MAP domain to a traffic bottleneck.

Other aspects of MIPv6, such as the multi-cast of router advertisements for mobility detection can be incorporated in the network discovery mechanism of the proposed architecture. Router advertisements are used in IPv6 and MIPv6 for stating the availability of a router. A mobile terminal receives router advertisements and forwards them to AF-agent in order to subscribe to NM-agent to network descriptions that correspond to the discovered access router. The new network descriptions are then utilized by AF-agent in its decision process. On the other hand, when router advertisements of the current network are no longer received by the MMT, AF-agent is notified in order to unsubscribe from the respective network descriptions.

4.5 Performance evaluation

In [CVS$^+$04] a partitioning of the handover latency is proposed, where the total latency T_{HO}, is broken into three main components, namely detection period T_d, address configuration interval T_c and network registration time T_r, $T_{HO} = T_d + T_c + T_r$. As the main focus of the proposed approach is on the selection of the timing and target for handover execution, the latency introduced to the handover procedure due to its operation corresponds to the detection period component of the above partitioning.

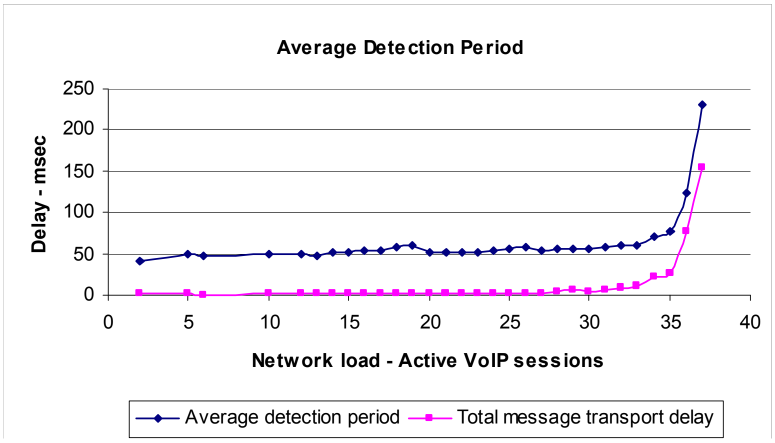

Figure 4.7: Average Detection Period on various network loads.

A simulation system has been implemented in order to estimate T_d and study the performance of agent migration. System implementation is based on JADE 3.4 [jad10a], a widely-used framework for developing agent-based applications. The simulation setting that has been deployed in order to estimate T_d comprises two agent platforms P_1 and P_2 that serve the users of two IEEE 802.11g Access Points (APs), AP_1 and AP_2 respectively. Each platform executes on a 3.2 GHz Pentium 4 workstation with 1 GB RAM and hosts the set of agents described in Section 4.3.3. As concerning user representative agents, an instance of AF-agent is executed for each MMT that is associated with an AP, regardless

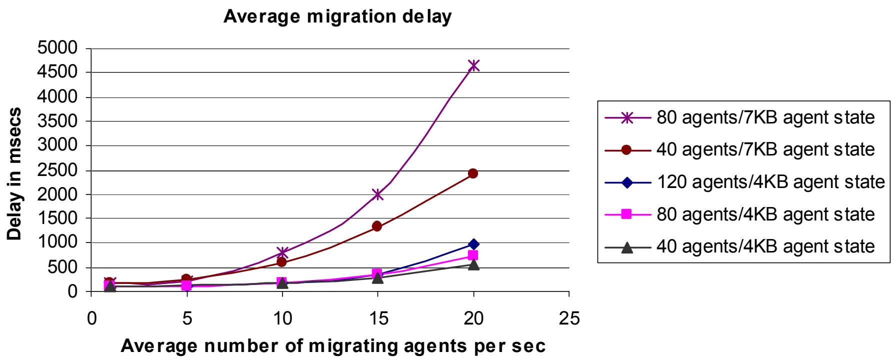

Figure 4.8: Average Migration Delay.

of having or not an active data session with it. A third agent platform P_0 hosts P-agents, that correspond one to one to AF-agents executing in P_1 and P_2. Each P-agent generates VoIP call establishment requests *REQ* that trigger handover decisions to their corresponding AF-agents. AF-agents incorporate an implementation of the AHP-GRA network selection algorithm [SJ05] that is executed upon each request. The result *RES* is sent back to P-agent and the round-trip time corresponds to T_d.

The VoIP service is selected as a type of traffic in order to study performance results in a setting with strict QoS requirements. The call duration and inter-arrival times are exponentially distributed. Call servicing and thus the network load is simulated by NP-agents. The average packet delay under various network loads is estimated by a WLAN simulator, Pamvotis [VZ10]. Pamvotis is also used for an estimate of the AP's capacity, in terms of VoIP sessions. As the VoIP service requires delay values to be less than 150ms in each direction, capacity threshold is set to a number of sessions that satisfies this requirement. For the simulation setting included in Table 4.2 the AP capacity is 37 sessions.

Table 4.2: Simulation parameters

Pythagor Simulation parameters		***Network load generation parameters***	
AP Data rate	11Mbps	**Call Duration**	Exponential μ=0,0083s^{-1}
VoIP Packet length	200B	**Call Arrival**	Poisson λ=0,245s^{-1}
VoIP Packet rate[2]	25pkt/s	**AP Capacity**	C=37 VoIP sessions
VoIP Bit rate	80Kbps	**Call Blocking Prob.**	30%

---

[2] One way packet rate. A total of 80Kbps is required for a single call, 40Kbps in each direction.

The average T$_d$ has been estimated for 200 AF-agents executing in each of P$_1$, P$_2$. The simulation setting involves high utilization of both WLAN APs with call blocking probability equal to 30%. The average duration of VoIP sessions is set to $d = 120s$ and, thus, random duration values follow an exponential distribution with parameter $\mu = 0,0083s^{-1}$. Given that VoIP call arrivals follow a Poisson distribution, the system can be modelled as M/M/m/m queueing system, i.e., a system with m=37 servers and no waiting room. The call blocking probability of this system is given by Erlang B loss formula [Kle75]

$$P_m = \frac{\frac{E^m}{m!}}{\sum_{i=0}^{m} \frac{E^i}{i!}},$$

that is used for estimation of λ, given the values of P_m, μ and $E = \frac{\lambda}{\mu}$. Figure 4.7 presents the average T$_d$ as it evolves with the number of active sessions on the AP. The message transport delay (T$_m$) corresponds to the latency introduced by the wireless access network for the transport of both *REQ* and *RES*. The average message size for REQ and RES is 300B and 1700B respectively. The diagram shows that T$_d$ is equal to T$_m$ plus an overhead of about 50ms due to agent collaboration. Thus, T$_d$ remains in reasonable levels, considering that in case of WLAN/GPRS vertical handovers it has values in the order of 100ms [CVS$^+$04]. In future extensions of this work, T$_d$ will be studied under larger scale deployments.

As JADE 3.4 does not support inter-platform mobility, a third-party service has been employed for the implementation of agent migration in the simulation system. The Inter-Platform Mobility Service (IPMS) [Dep10] integrates with the JADE agent platform and utilizes the platform's message transport service for implementing agent migration between different JADE platforms. In IPMS, the platform's AMS packages the migrating agent's code and state in an ACL message - ACL$_m$ - and sends it to its peer AMS in the remote platform that restores the agent's execution.

A simulation setting comprising two JADE platforms has been deployed in order to study the performance of agent migration. Each platform executes a number N of AF-agents that migrate at exponentially distributed intervals, thus, generating a Poisson distributed series of migrations with a total rate λ. Experiments have been performed for various values of λ and N and the results are presented in Figure 4.8. A third determinant of migration delay is the size of the migrating agent's code and state. Due to the relatively large size of AF-agent's code (100KB) a modification was introduced to the IPMS source code in order to disable agent code packaging in ACL$_m$. The assumption is that agent code will be pro-actively transferred and cached to platforms of “neighbouring” cells before handover execution. The graph in Figure 4.8 shows a significant increase in migration latency (especially for high migration rates) as the agent state raises from 4KB to 7KB. A good design of AF-agents that limits the size of agent state, as well as the use of high processing power servers for hosting the agent platforms can moderate the migration delay. However, it must be highlighted that migration delay does not contribute to handover latency, as agent migration takes place in parallel with handover execution. The restriction here is that its value should be lower than handover latency so that AF-agent will be available to the MMT when the connection to the new network is established.

Chapter 5

Conclusions & Future work

5.1 Conclusions

This thesis contributes to the broader area of network selection and handover decision in a heterogeneous network setting. This network setting, ofter referred to as 4G, is characterized by multiple Radio Access Networks (RANs), of possibly different Radio Access Technology (RAT), interworking through a common IP core network or different core networks interconnected through the Internet infrastructure. Network selection is part of the handover management procedure and targets the selection of the best point of access and service level for maintaining the level of connectivity required by user applications. In a heterogeneous network setting the range of choices allows network selection to serve additional user objectives to service continuity, e.g., economic efficiency, device energy autonomy. This thesis studies the network selection problem in the presence of multi-homing support from both the network side and the end-host. Specifically, a mobile terminal is assumed, equipped with two or more different radio interfaces that has the role of an end-host or mobile router. The multi-homing capability widens the scope of network selection to also involve the selection of radio interfaces that need to be activated and the decision on the distribution of traffic flows to them. The latter is necessary since traffic requirements are an important decision factor on the majority of network selection schemes. Thus, the problem has three dimensions: (a) radio interface activation, (b) bearer service selection for activated radio interfaces, (c) assignment of traffic flows to activated radio interfaces.

In this thesis the three subproblems are handled in a uniform manner with the specification of the Traffic Flow Assignment Problem (TFAP). TFAP focuses on the assignment of application traffic flows (either inbound or outbound) on appropriate radio interfaces and bearer services in a way that establishes the best trade-off between economic cost and power consumption. Economic cost refers to network usage cost while power consumption is due to the operation of activated radio interfaces. The assumptions that underpin the trade-off between them are: (a) the availability of bearer services of comparable cost across the mobile terminal's radio interfaces (b) the existence of a positive correlation between network usage cost and provided capacity. Based on these assumptions and given that the traffic load of the Mobile Multi-mode Terminal (MMT) cannot be served by a single radio interface (due to capacity or QoS restrictions), the minimization of economic cost involves distributing traffic flows to radio interfaces with access to the cheapest bearer services. On the other hand, minimization of power consumption requires the activation of the least possible number of radio interfaces by associating them with high capacity and usually higher cost bearer services

The proposed analytic formulation for TFAP results in a bi-objective combinatorial optimization problem. The bi-objective optimization problem is solved after its transformation to a single-objective optimization problem. Specifically, economic cost is selected as a primary objective for optimization, while power consumption is used as an additional problem constraint through the definition of an upper limit for its allowed values. The limit on power consumption is not constant for all problem instances but depends on device status (e.g., energy reserves) and context and is calculated, when required, by the mobile terminal's power management subsystem. This thesis includes a study on TFAP's complexity through reduction from the Multiple Knapsack problem with Assignment Restrictions (MKAR). Since MKAR is NP-hard, TFAP is also NP-hard and approximation algorithms are required for fast derivation of problem solutions.

A heuristic local search algorithm is introduced towards this direction that is characterized by efficient execution times for a wide set of realistic problem sizes. The quality of approximation is rather satisfactory and is evaluated through comparison of heuristic and exact solutions for a large set of randomly generated problem instances. Specifically, the algorithm's approximation error in economic cost has an average value below 8.1% for problems of small to medium size. Regarding the distribution of approximation error, 80% of these problem instances are solved with approximation error lower than 15%, while 95% of them are solved with error lower than 35%. Since obtaining exact solutions for large problem instances is a rather time-consuming procedure, the evaluation of algorithm's quality of solutions against such problems is based on exact solutions of a relaxation of TFAP. The relaxed problem derives from TFAP by removing the flow integrality constraint, i.e., by allowing the traffic of each individual flow to be split across two or more radio interfaces. Its solutions are better or equal to solutions of the original problem and thus the approximation error is overestimated. Despite this fact, the average approximation error over the set of solved large problem instances is below 13.1%. With regard to the distribution of error, 80% of large problems are solved with accuracy higher than 20%, while 90% of them with higher than 29%.

The merits of optimized traffic flow assignment when applied over a specific time horizon, as well as the associated mobility management overhead has been evaluated through simulation. The implemented discrete event simulator simulates a three hour operation of a mobile router serving 5 users and equipped with 2 UMTS and 2 WLAN radio interfaces. UMTS interfaces have access to 4 bearer services, while each WLAN interface has access to 10 bearer services. Each user generates video-conference and FTP/HTTP sessions with arrival rate and duration that follow widely accepted traffic models. The system also simulates the adaptation of the mobile router to changing network and traffic conditions through iteratively solving TFAP problem instances. Averaged results over 100 simulation executions show a 7.5% increase of the total economic cost against optimal cost, when applying the proposed heuristic algorithm. On the other hand, the employment of an alternative algorithm proposed in the literature results to 17% cost increase. The mobility management overhead due to enforcing TFAP solutions involves an average of 2.2 flow redirections and 1.5 horizontal handovers per minute, that is tolerable given the number of served users. The economic cost savings combined with the limited mobility management overhead allow a practical deployment of the proposed heuristic algorithm.

The employment of advanced network selection (or handover decision) or TFAP algorithms cannot be based solely on an end-host infrastructure. The decision mechanisms require both locally available information (e.g., application traffic requirements, user preferences) and location-based network information that is not practical to be retrieved by the mobile terminal exclusively through active scanning. The reason is that active scanning is time consuming and inefficient in terms of energy consumption. Moreover, reliable and in-time information on resource availability of available access networks may not be provided by a single network operator, e.g., the home operator of a mobile user. This is due to its lack of incentives for providing it and consequently letting its clients utilize third-party services.

In this thesis a system architecture is proposed for supporting the execution of handover decisions or TFAP algorithms. The architecture spans multiple administrative domains and is based on software agents that execute in agent platforms deployed in these domains. The software agents represent the users, the network operators, a Multi-Access Provider (MAP) and the regulatory authority. MAP is a business entity that maintains roaming agreements with network operators and enables user utilization of their services through a single subscription. Moreover, MAP provides AAA, billing and inter-domain mobility management support. The regulator enhances user trust by monitoring the behavior of the operators and intervening when required. In the proposed approach, handover or traffic flow assignment decisions (in single-homed or multi-homed hosts respectively) are initiated by user agents that execute either in the terminal or the network side, depending on the source of triggering events. Decision making is delegated to software agents that execute in the network side for saving the terminal's usually limited power and computational resources. These agents also employ agent mobility in order to migrate and execute to platforms that are "closer" in terms of network delay to their corresponding mobile terminals. Thus, mobile terminal responsiveness to decision triggering events is enhanced. The thesis proposes the selection of both access and core bearer services during decision making for better adaptation to user and application requirements. Towards this purpose, a bearer service data model is presented, as well as a procedure for utilizing bearer service descriptions in handover or TFAP decision algorithms.

Performance evaluation of the proposed architecture has been performed through a simulation system implemented in Java and based on the JADE framework for agent-based applications. The focus of the simulation is twofold: (a) estimation of the latency introduced to the handover detection period due to agent collaboration, (b) study of the impact of agent mobility on the proposed system's performance. Simulation results show a limited overhead, in the order of 50ms, to the handover detection period that does not severely impact mobile terminal's responsiveness to handover triggering events. The study of agent mobility under different rates of agents, migrating between two agent platforms, identified the agent state size as a determining performance factor (in addition to agents' migration rate). Specifically, the larger the size of the mobile agent's state, the higher the time required to transfer and restore its execution state in the target agent platform (migration delay). On the basis of simulation results, an agent state of 4KB or lower allows high rates of migrating agents

(120 agents/sec) while keeping agent migration delay lower than 1sec. Note that agent migration does not affect handover execution (it takes place concurrently with it) and user agent execution in the target platform must be restored in order to enable subsequent handover decisions. Thus, agent migration delay does not degrade system's responsiveness to forthcoming handover triggering events.

5.2 Future Work

Future work will involve extensions to TFAP problem formulation, as well as to the respective flow assignment algorithm so as to handle special requirements such as: (a) support of real-time flows with alternative levels of bandwidth requirements, where each level may correspond to a different codec or codec configuration, (b) support of flows that use transport protocols with Concurrent Multi-path Transfer capabilities and can be distributed to two or more radio interfaces. Moreover, the proposed traffic flow assignment scheme will be enhanced with a decision mechanism for automatically fixing the limit on power consumption in response to changes in served traffic, minimum energy autonomy preferences, battery charging level etc. Last but not least, future research will embrace issues related to inferring the application type or QoS requirements of traffic flows in cases that this information cannot be obtained from the execution environment of the flow assignment algorithm. This capability is more important for a mobile node that has the role of an Internet gateway in a personal area or vehicular network.

The future work will also focus on architecture enhancements and evaluation of its efficiency on the basis of a system prototype embedded in real a wireless networking environment. Joint resource and handover management performed by agents on network providers' platforms for global load balancing will also be studied.

Appendices

Appendix A

Implementation details on the evaluation of the TFAP heuristic

This appendix provides details on a Java implementation of the TFAP heuristic local search algorithm introduced in Section 3.3. Moreover, it describes the implementation of the systems that were used for the TFAP heuristic algorithm evaluation of Section 3.4.

A.1 Implementation and evaluation of TFAP heuristic

A core part of the TFAP heuristic algorithm implementation is the `TFAProblem` class that models a traffic flow assignment problem, as defined in Section 3.2.1. Figure A.1 presents the structure of `TFAProblem`, as well as relationships among the rest of the classes that model the TFAP problem domain. Specifically, `Flow`, `Bearer` and `RadioInterface` classes represent the respective concepts introduced in Section 3.2.1. A `TFAProblem` instance incorporates all flows, bearer services and radio interfaces that define a TFAP instance.

A traffic flow assignment is represented by the state of `RadioInterface` class instances. The state of a `RadioInterface` instance comprises (a) a list of `Flow` instances that correspond to flows that are served by the radio interface, and (b) a `Bearer` instance that represents the bearer service that is associated with it. The `associatedBearer` attribute takes its value from a list of bearer services that are compatible with the radio interface (`compatibleBearers` attribute). Note that the value of `associatedBearer` is `null` in case that the radio interface is deactivated.

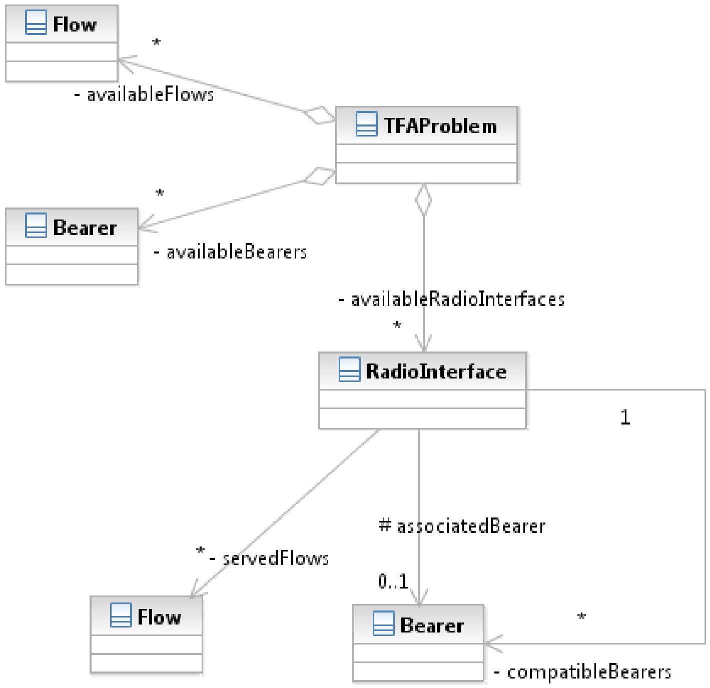

Figure A.1: Traffic flow assignment domain model.

On the basis of `RadioInterface` instances' state, `TFAProblem` class provides methods for returning the economic cost and power consumption of its current traffic flow assignment.

The design of the TFAP heuristic local search algorithm is illustrated in Figure A.2. The algorithm is implemented by `HeuristicScheduler` class that provides the `IScheduler` interface. This interface is also supported by the other flow assignment algorithm implementations that are used for the evaluation of the proposed algorithm. `HeuristicScheduler` uses the factories `ConstrHeuristicFactory` and `ObjFunctionFactory` for obtaining instances of "construction" heuristic algorithms and objective functions, respectively, that are required for its operation as specified in Section 3.3. Specifically, `FirstFitWRAlgorithm` implements Algorithm 4 and its variation for constructing a minimum power consumption solution. Moreover, `CostOptMinPwrIncrFunction` and `PwrConsOptMinCostIncrFunction` instances are used for evaluating alternative problem states on the basis of functions f and g respectively.

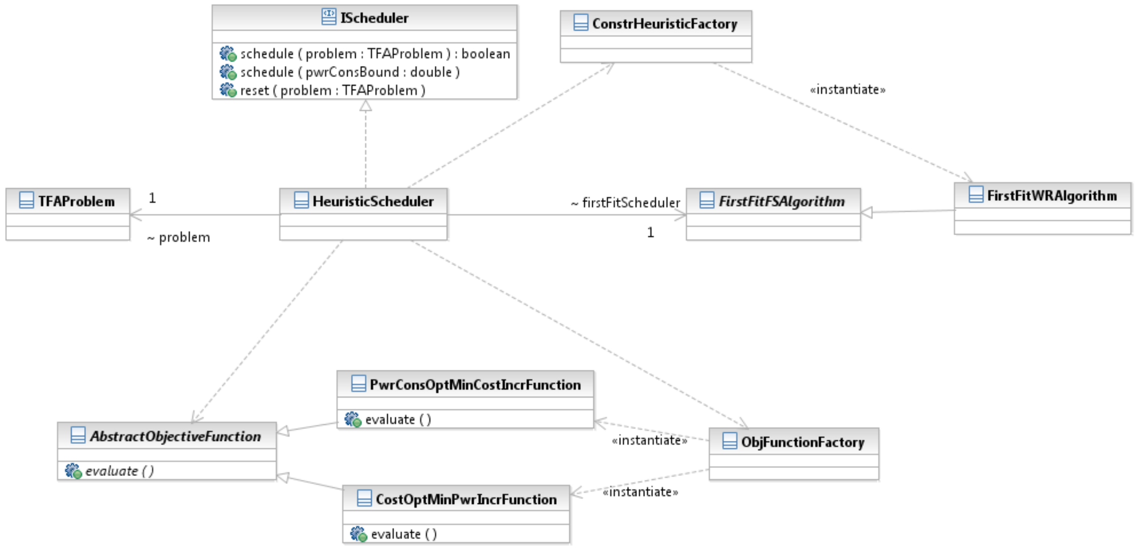

Figure A.2: TFAP heuristic local search algorithm implementation.

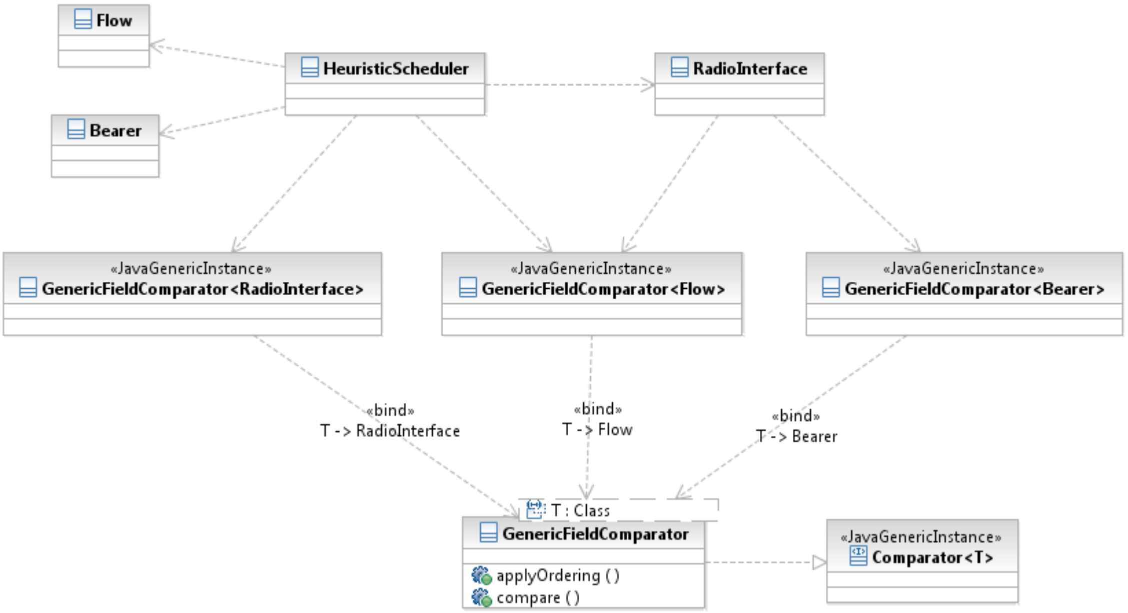

Figure A.3: HeuristicScheduler dependencies on domain model classes.

Figure A.3 depicts the dependencies of `HeuristicScheduler` on domain classes. As specified in Section 3.3, the algorithm operates on the state of `RadioInterface` instances of a `TFAProblem` instance, causing, thus, problem state transitions. The various sorting operations specified by Algorithms 2, 3 (Section 3.3) are supported by instances of the `GenericFieldComparator` generic class. The class allows declarative definition of comparators based on attributes of the class that has been provided as template parameter.

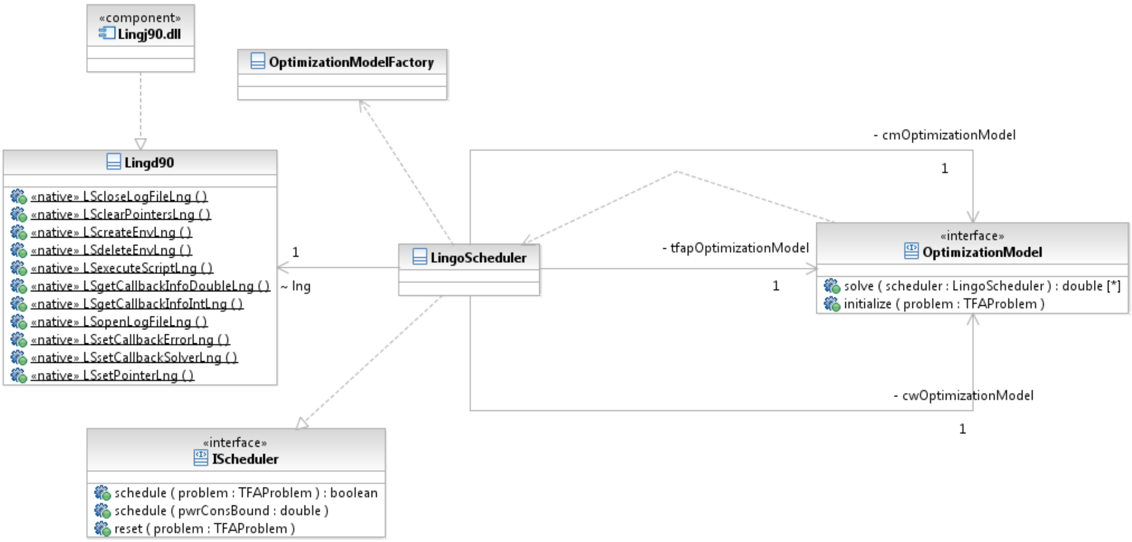
«component»
Lingj90.dll
OptimizationModelFactory
Lingd90
«native» LScloseLogFileLng ()
«native» LSclearPointersLng ()
«native» LScreateEnvLng ()
«native» LSdeleteEnvLng ()
«native» LSexecuteScriptLng ()
«native» LSgetCallbackInfoDoubleLng ()
«native» LSgetCallbackInfoIntLng ()
«native» LSopenLogFileLng ()
«native» LSsetCallbackErrorLng ()
«native» LSsetCallbackSolverLng ()
«native» LSsetPointerLng ()
1
~ lng
LingoScheduler
- cmOptimizationModel
1
- tfapOptimizationModel
1
«interface»
OptimizationModel
solve (scheduler : LingoScheduler) : double [*]
initialize (problem : TFAProblem)
- cwOptimizationModel
1
«interface»
IScheduler
schedule (problem : TFAProblem) : boolean
schedule (pwrConsBound : double)
reset (problem : TFAProblem)

Figure A.4: LingoScheduler design.

Exact problem solutions are obtained through `LingoScheduler` that also implements the `IScheduler` interface. `LingoScheduler` wraps the Lingo 9.0 optimization library and invokes its operations through Java Native Interface (JNI). A `LingoScheduler` instance creates the Lingo execution environment and pointers to arrays of `double` type that (a) store problem parameters to be passed to Lingo and (b) act as placeholders for problem solutions. `LingoScheduler` obtains the solution of a problem instance through a native method invocation that takes as parameter a script with solver configuration properties and the path to a Lingo script with the problem formulation. Figure A.4 presents the design of `LingoScheduler` that will be further explained hereafter.

Note that `LingoScheduler` is required to solve different variations of the TFAP problem, i.e., TFAP, TFAP$_M$ and TFAP$_P$ with reference to Section 3.2.2. Each variation is solved by a different Lingo script and passes different arguments to the Lingo environment. In order to make the implementation of `LingoScheduler` more generic, the requirements set by each problem type, in terms of Lingo arguments and solution script,

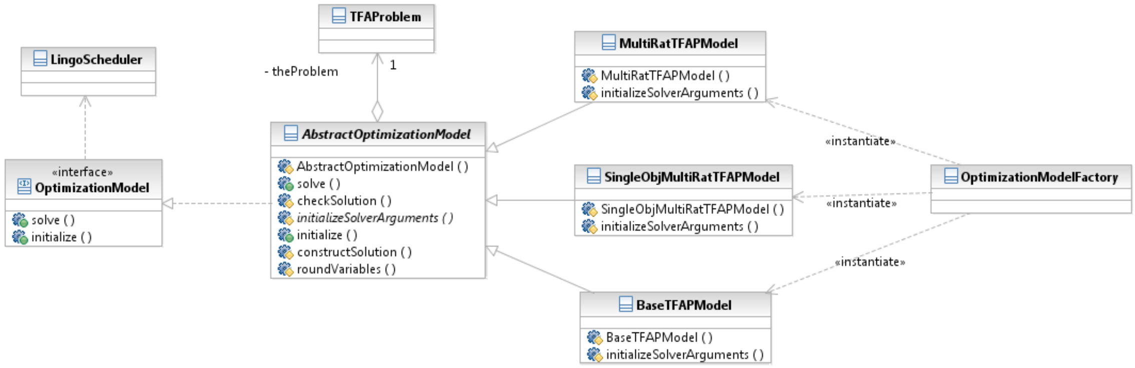

Figure A.5: Detailed design of LingoScheduler.

have been encapsulated in implementations of the `OptimizationModel` interface. Each instance conforming to that interface is initialized with the `TFAProblem` instance and extracts from it the required data to be passed as Lingo arguments. Upon invocation of the `solve` method, the `OptimizationModel` instance uses the `LingoScheduler` interface to provide the required arguments to the Lingo environment and trigger its execution with an appropriate script with the problem formulation. Figure A.4 depicts the dependencies of `LingoScheduler` on `OptimizationModel` instances. Specifically, it has references to three instances, `tfapOptimizationModel`, `cmOptimizationModel` and `cwOptimizationModel`, corresponding to $TFAP$, $TFAP_M$ and $TFAP_P$ problems respectively. The `tfapOptimizationModel` is an instance of `MultiRatTFAPModel` class while the other two of the `SingleObjMultiRatTFAPModel` class (see Figure A.5).

Figure A.6 presents the utility classes that were used for generating and solving the random problem instances required for the evaluation of the heuristic algorithm. Class `TFAPGenerator` generates different TFAP problem instances, finds their exact solutions (through `LingoScheduler`) for different values of the limit on power consumption and stores problem definitions and solutions in a data-store. On the other hand, `TFAPBatch-Solver` retrieves problem definitions, solves them with a flow assignment algorithm and stores their solutions in a data-store. Both classes depend on the `ProblemGenerator` and `ProblemSink` interfaces. `ProblemGenerator` specifies an operation for retrieving TFAP problem instances. Depending on the implementation, problem instances might be randomly generated (`RandomDiscrTFAPGenerator`) or retrieved from a relational database (`ProblemManager`). The interface `ProblemSink` defines operations for storing

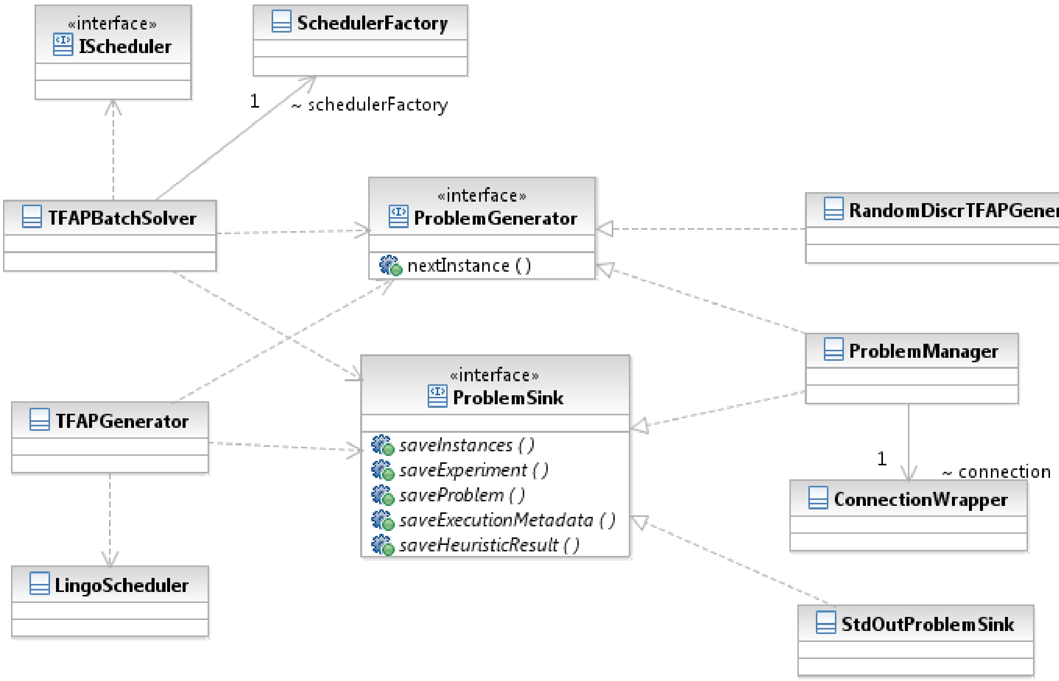

Figure A.6: Problem generation and solving utilities.

TFAP problem definitions and their solutions to a datastore. The `ProblemManager` implementation stores them to a relational database while `StdOutProblemSink` justs prints them to the standard output for debugging purposes. During the random problem generation process, `ProblemGenerator` and `ProblemSink` interfaces of `TFAPGenerator` are bound to `RandomDiscrTFAPGenerator` and `ProblemManager` classes respectively. On the other hand, during batch solving of the generated problem instances with a heuristic algorithm the aforementioned interfaces are both bound to `ProblemManager`. With regard to the `IScheduler` interface, it is bound to either one of `HeuristicScheduler` and `UtilityScheduler` classes (see Figure A.7).

Figure A.7 presents the different types of flow assignment algorithms used for evaluation of the proposed TFAP heuristic. Class `UtilityScheduler` implements the utility-based flow assignment algorithm described in Section 3.4.2 and initially proposed in [NVAGD08].

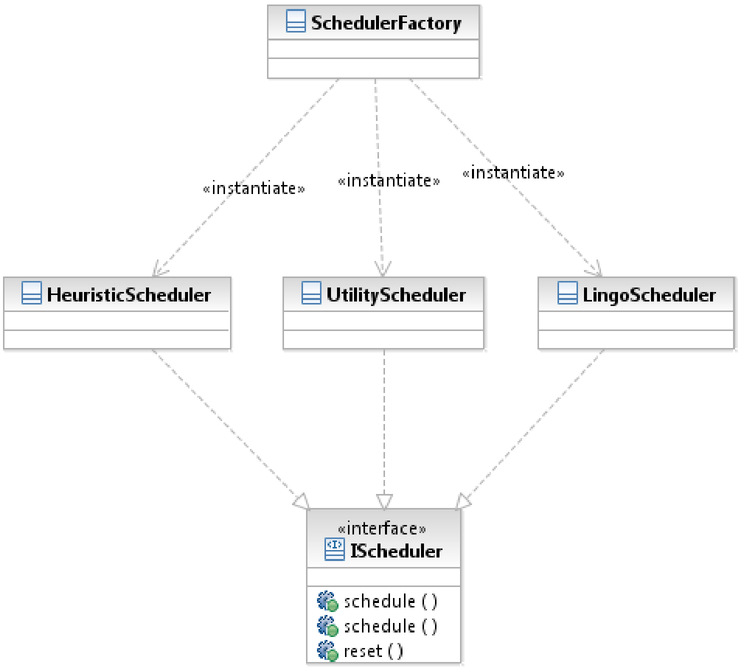

Figure A.7: Flow assignment algorithm implementations used in the evaluation.

A.2 TFAP heuristic evaluation through simulation

In this section a description of the implementation of the simulation system used for the evaluation of Section 3.4.2 will be provided. The top level components of the simulator are represented by `CompositeEventSource`, `DecisionEngine` and `MobileHost` classes. A `MobileHost` instance, that represents the simulated mobile router/host, is created upon each simulation execution. This instance creates and configures one instance from each one of `CompositeEventSource`, `DecisionEngine` and controls the progress of the simulation. Specifically, `MobileHost` retrieves sequentially events from the `CompositeEventSource` and passes them to the `DecisionEngine` in order to be processed. Figure A.8 presents the structure of `MobileHost` and `DecisionEngine` classes.

The `DecisionEngine` holds an instance of `TFAProblem` (`problem`) that represents the current flow assignment status throughout the simulation duration. `DecisionEngine` updates `problem` state on the basis of event arrivals. The types of events that receives from its execution context are network and flow related events that are instances of the

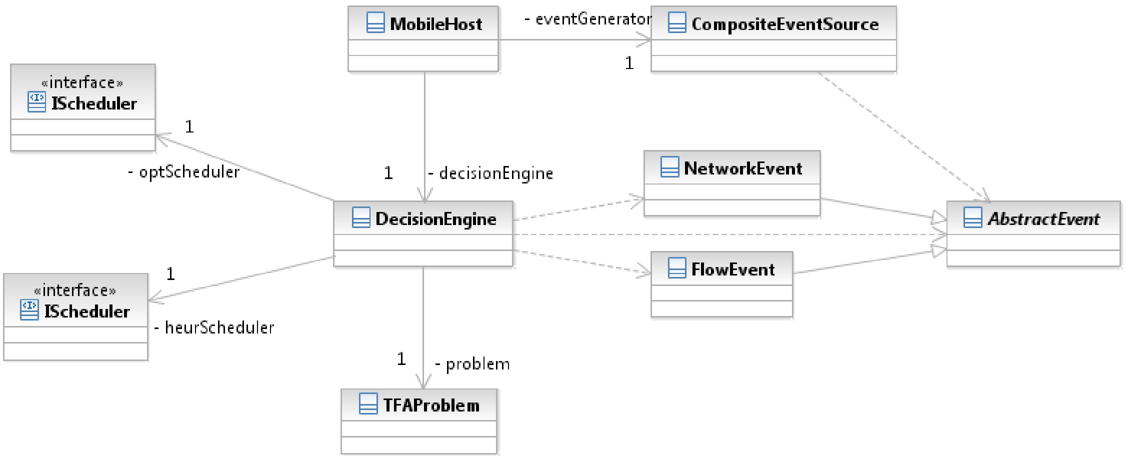

Figure A.8: Top level design of the simulator.

`NetworkEvent` and `FlowEvent` classes respectively. Network event types correspond to (a) discovery of a new RAN, (b) unavailability of a RAN and (c) change to the status of an available RAN. On the other hand, flow events correspond to arrival or termination of a set of traffic flows. Depending on the type of a `NetworkEvent` or `FlowEvent` instance that `DecisionEngine` processes, certain operations are applied to its `problem` attribute:

- network discovery → append the bearer services included in the `NetworkEvent` instance to the available bearers of `problem`,
- network unavailability → remove the bearer services corresponding to the RAN that the `NetworkEvent` refers to from the `problem`'s state,
- network status change → remove from `problem` the bearer services corresponding to the RAN that the `NetworkEvent` refers to and append the bearers included in the event instance,
- flow arrival → append the flows included in the `FlowEvent` instance to the available flows of `problem`,
- flow termination → remove the flows included in the `FlowEvent` instance from the `problem`'s state.

The change to the `problem`'s state due to the arrival of an event is followed by the execution of the heuristic (either one of `HeuristicScheduler` or `UtilityScheduler`) and the exact (`LingoScheduler`) flow assignment algorithms. The algorithms use the `problem` instance as input data and their execution results to probably different flow assignments. The economic cost and power consumption of each assignment is used to update the total cost and power consumption of each approach for the time duration until the arrival of the next event.

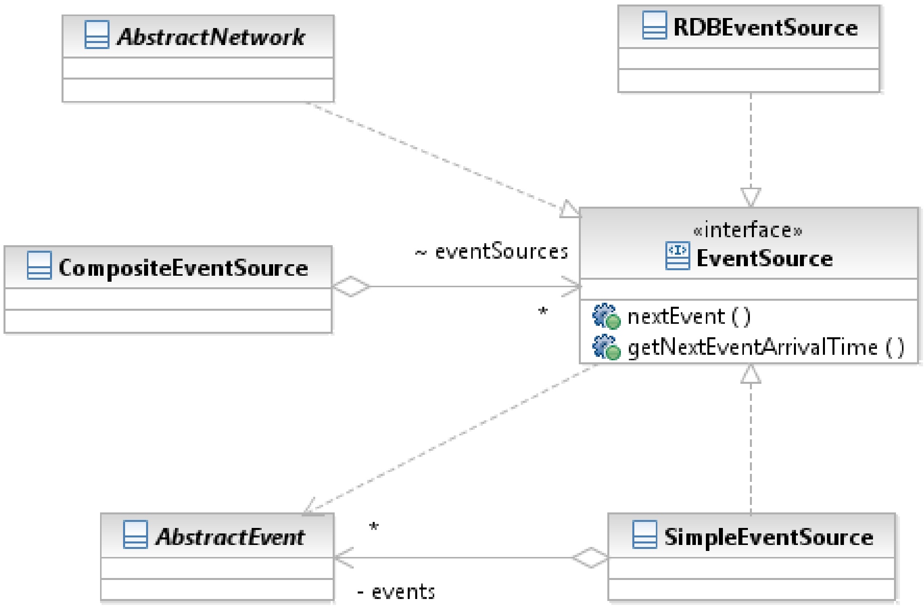

Figure A.9: Classes implementing event sources.

The `CompositeEventSource` comprises a set of event sources that implement the `EventSource` interface (see Figure A.9). The interface specifies an event queue that is manipulated through the `nextEvent` operation. The latter removes an event from the queue, while `getNextEventArrivalTime` operation returns the timestamp that corresponds to the arrival of the next event. `CompositeEventSource` manages multiple `EventSource` instances by peeking each time an event from the source with the smallest arrival time. The `AbstractNetwork` implementation of `EventSource` corresponds to a simulated RAN and generates `NetworkEvent` instances. The `SimpleEventSource` implements a simple event queue that does not generate events but instead is initialized upon its creation with a set of events. This event source is used to feed the simulator with flow events. A typical runtime configuration of `CompositeEventSource` comprises one instance of `SimpleEventSource` and multiple instances of `AbstractNetwork`.

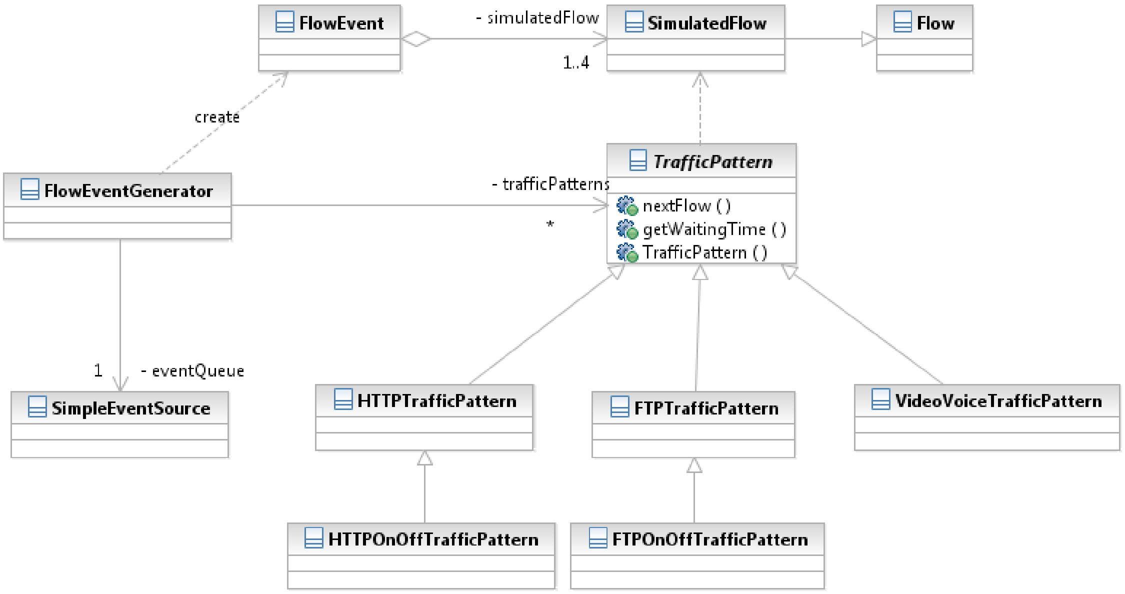

Figure A.10: Flow event generation infrastructure.

Figure A.10 illustrates the design of `FlowEventGenerator` class that is responsible for initialization of a `SimpleEventSource` instance with all flow events that will be used in a simulation. The `FlowEventGenerator` is initialized with a set of `TrafficPattern` instances that generate the simulated traffic flows. Each simulated traffic flow is characterized by QoS class, bandwidth and duration attributes. The `TrafficPattern`'s `nextFlow` operation returns a set of 1, 2 or 4 flows depending on the `TrafficPattern` implementation. Specifically, non real-time traffic patterns always return a single flow, while instances of `VideoVoiceTrafficPattern` return 2 or 4 flows depending on whether they are configured to generate VoIP or video-conference sessions. The `getWaitingTime` operation returns the time interval in ms until the arrival of the next flow. Each `TrafficPattern` instance models the activity of a single user with respect to a certain type of service, e.g., web browsing, ftp transfers, video/voice calls and so on.

The `FlowEventGenerator` processes sequentially each `TrafficPattern` instance for a total simulated time equal to the simulation duration that is given as system parameter. Each call to `nextFlow` is followed by the creation of a pair of flow arrival and flow termination events that concern the generated flow(s). The listing below describes in Java the generation of events due to the processing of a single `TrafficPattern` instance:

Listing A.1: Flow event generation from a single `TrafficPattern`

```
protected List<FlowEvent> generateEventsForTrafficPattern(TrafficPattern tp,
    long duration) {

  // Simulation duration in milliseconds
  long durationMillis = 1000 * duration;
  long currentTime = 0;
  ArrayList<FlowEvent> eventList = new ArrayList<FlowEvent>();
  FlowEvent startEvent, stopEvent;
  List<SimulatedFlow> flows = null;
  SimulatedFlow f;

  // update current time with waiting time till first flow arrival
  currentTime += tp.getWaitingTime();
  while (currentTime < durationMillis) {
    // get the newly arrived flows
    flows = tp.nextFlow();

    // create the flow start event
    startEvent = new FlowEvent(currentTime, flows, EVENT_TYPE.FLOW_START, tp.
        getId());
    eventList.add(startEvent);

    f = flows.get(0);
    // update current time with flow duration
    currentTime += f.getDuration();

    // create the flow termination event
    stopEvent = new FlowEvent(currentTime, flows, EVENT_TYPE.FLOW_TERM, tp.
        getId());
    eventList.add(stopEvent);

    // update current time with the waiting time until next flow
    currentTime += tp.getWaitingTime();
  }
```

```
    return eventList;
  }
```

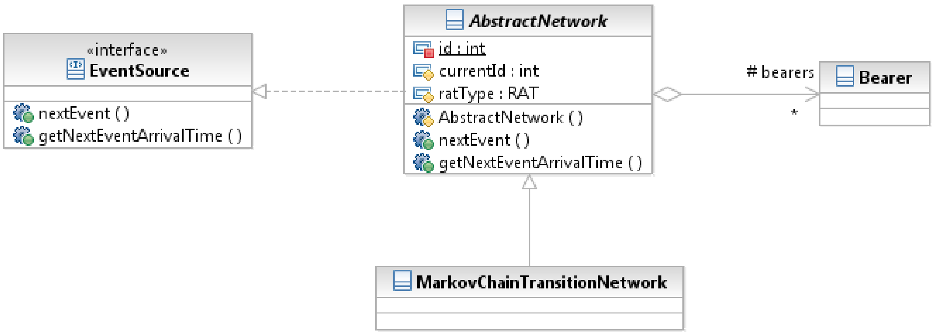

Figure A.11: Network events' source design.

Finally, Figure A.11 presents the design of `MarkovChainTransitionNetwork` class that implements a simulated RAN used in the simulation system.

A.3 TFAP problem formulation in Lingo

This section includes the Integer Linear Programming (ILP) formulation of the TFAP problem expressed in the syntax of the Lingo environment for solving optimization problems. The script included in the following listing is wrapped by `MultiRatTFAPModel` instances used by the `LingoScheduler` for producing exact solutions of TFAP instances. The explanatory comments related to this script are included in the listing as Lingo comments (beginning with "!"). Moreover, assignment statements that have a "@POINTER(X)" expression on the right side of the assignment operator correspond to variable initializations with script input data. On the other hand, statements with a "@POINTER(X)" expression on the left side of the assignment operator correspond to returned values by the script to its calling context.

Listing A.2: TFAP ILP formulation in Lingo

```
MODEL:
DATA :
```

```
! Input data ;
 BOUND = @POINTER(1); ! Limit on power consumption
 NUM_IFS = @POINTER(2); ! Number of radio interfaces
 NUM_BEARERS = @POINTER(3); ! Number of bearer services
 NUM_FLOWS = @POINTER(4); ! Number of traffic flows
ENDDATA
SETS:
! Input data regarding bearer services are sorted by compatibility ;
! with radio interfaces, i.e., first the elements of B_1, then of ;
! B_2, and last of B_m ;
! The following set defines the start and end index of each ;
! subset B_1, B_2, ... B_m in the set of bearer services ;
 RI_TO_BEARER_MAP /1.. NUM_IFS/ :FROM, TO;

! Radio interfaces ;
 INTERFACE / 1..NUM_IFS / ;
! Bearer services and their attributes ;
 NETWORK /1..NUM_BEARERS / : UP_BW, DOWN_BW, DELAY, COST, POWER_CONS_IDLE,
     POWER_CONS_TR, POWER_CONS_RCV;
! Flows and their attributes. Direction value of 1 denotes ;
! an upstream flow, while a value of 0 a downstream one;
 FLOW /1..NUM_FLOWS / : DIRECTION, MAX_DELAY, BANDWIDTH;

! Valid combinations of interfaces and bearer services. Each ;
! combination is characterized by a binary variable Y ;
 NETWORK_SELECTION_ALT (INTERFACE,NETWORK) | &2 #LE# TO(&1) #AND# &2 #GE#
     FROM(&1) :Y;
! Combinations of bearer services and flows. Each combination  ;
!is characterized by a binary variable X ;
 FLOW_ASSIGNMENT(NETWORK_SELECTION_ALT, FLOW):X;
ENDSETS

W_R = @SUM(FLOW_ASSIGNMENT(I,J,Z):POWER_CONS_RCV(J)*X(I,J,Z)*(1 - DIRECTION(Z
    ))*BANDWIDTH(Z));

W_T = @SUM(FLOW_ASSIGNMENT(I,J,Z):POWER_CONS_TR(J)*X(I,J,Z)*DIRECTION(Z)*
    BANDWIDTH(Z));

W_I = @SUM(NETWORK_SELECTION_ALT(I,J):POWER_CONS_IDLE(J)*Y(I,J));
! Calculation of the economic cost of a flow assignment ;
C_M = @SUM(FLOW_ASSIGNMENT(I,J,Z):COST(J)*X(I,J,Z)*BANDWIDTH(Z));
```

```
! Calculation of the power consumption of a flow assignment ;
C_W = W_R + W_T + W_I;

! Objective is the minimization of C_M that as specified above ;
[OBJECTIVE] MIN = C_M;

! Problem constraints are described below ;
! Constraint on power consumption ;
C_W - BOUND <= 0;

! Association of each radio interface with at most one bearer;
@FOR(INTERFACE(I):@SUM(NETWORK_SELECTION_ALT(I,J):Y(I,J))<=1);

! Assignment of each flow to at most one radio interface ;
@FOR(FLOW(Z):@SUM(FLOW_ASSIGNMENT(I,J,Z):X(I,J,Z))=1);

! Uplink capacity of used bearer services must not be violated;
@FOR(NETWORK_SELECTION_ALT(I,J):@SUM(FLOW(Z):BANDWIDTH(Z)*X(I,J,Z)*DIRECTION(
    Z)) - UP_BW(J)*Y(I,J) <= 0);

! Downlink capacity of used bearer services must not be violated;
@FOR(NETWORK_SELECTION_ALT(I,J):@SUM(FLOW(Z):BANDWIDTH(Z)*X(I,J,Z)*(1-
    DIRECTION(Z))) - DOWN_BW(J)*Y(I,J) <= 0);

! Flows’ maximum delay requirements must be met by serving bearer;
@FOR(FLOW(Z):@SUM(FLOW_ASSIGNMENT(I,J,Z):DELAY(J)*X(I,J,Z)) - MAX_DELAY(Z)<=
    0);

! Binary X variables;
@FOR(FLOW_ASSIGNMENT(I,J,Z):@BIN(X(I,J,Z)));

! Binary Y variables;
@FOR(NETWORK_SELECTION_ALT(I,J):@BIN(Y(I,J)));

DATA:
! Input data (continued from the beggining of the script)
! Arrays of attribute values for all bearer services;
POWER_CONS_IDLE = @POINTER(5); ! Base power consumption  ;
POWER_CONS_TR = @POINTER(6); ! Power consumption due to data transmission;
POWER_CONS_RCV = @POINTER(7);! Power consumption due to data reception;
```

```
UP_BW = @POINTER(8); ! Uplink bandwidth capacity;
DOWN_BW = @POINTER(9); ! Downlink bandwidth capacity;
DELAY = @POINTER(10); ! Maximum packet access delay;
COST = @POINTER(11); ! Economic cost;

! Arrays of attributes for all flows;
DIRECTION = @POINTER(12); ! Flow direction;
MAX_DELAY = @POINTER(13); ! Maximum delay;
BANDWIDTH = @POINTER(14); ! Required bandwidth capacity;

! Starting indices of sets B_1, B_2, ... B_m in the input ;
! arrays of bearer services attributes;
FROM = @POINTER(15);
! Ending indices of the aforementioned sets ;
TO = @POINTER(16);

! Output data
! Values of Y variables ;
@POINTER(17) = Y;
! Values of X variables ;
@POINTER(18) = X;
! Objective value (minimum economic cost) ;
@POINTER(19) = OBJECTIVE;
! Corresponding power consumption of solution;
@POINTER(20) = C_W;
! Problem solution status (feasible or not) ;
@POINTER(21) = @STATUS();
ENDDATA
END
```

Appendix B

Agent-based architecture evaluation

B.1 Simulation system implementation on JADE

This section presents design and implementation details of the simulation system used in evaluating the agent-based approach to handover decision support. The system is implemented in Java and based on JADE [jad10a]. JADE implements an agent platform that conforms to IEEE FIPA specifications [FIP10] and also provides a software infrastructure for developing agent-based applications. The JADE agent platform instantiates a container for execution of agent implementations and may be distributed across multiple Java Virtual Machines (JVMs) executing in the same or multiple hosts (even in mobile hosts). Any object inheriting from `jade.core.Agent` class is a software agent and, thus, may be hosted for execution in the JADE container.

The diagrams that follow illustrate the design of the agents specified in Section 4.3.3. The classes presented with their full qualified name belong to the JADE framework, while the rest of them have been implemented for the requirements of this thesis. The software agents that constitute the proposed architecture inherit indirectly from `jade.core.Agent` class through `BaseAgent`. The latter includes member variables common to all agents of the architecture, as well as utility methods required by them. Figure B.1 presents the design of both `BaseAgent` and `PAgent`. `BaseAgent` incorporates an instance of the ontology that defines the vocabulary of the problem domain (`HOOntology`). The `HOOntology` instance incorporates references to all classes that define the concepts, actions and predicates of the simulation system's problem domain. The definition of these classes is part of the system implementation and they realize respectively the `Concept`, `AgentAction` and `Predicate` interfaces that are defined in the `jade.content` package of the JADE framework. The `Codec` attribute represents an instance of a codec for conversion between object and string representations of information. The string representations follow the syntax of the FIPA-SL [FIP02d] content language and are assigned as values to the content slot of ACL messages. The validity of information during conversion is checked by the `Ontology` instance. The `Codec` and `Ontology` attributes of an agent are registered to a `jade.content.ContentManager` instance that is provided to each agent by the JADE container. The `ContentManager` provides an interface to agents for filling the content slot of ACL messages through object representations of the content and vice versa.

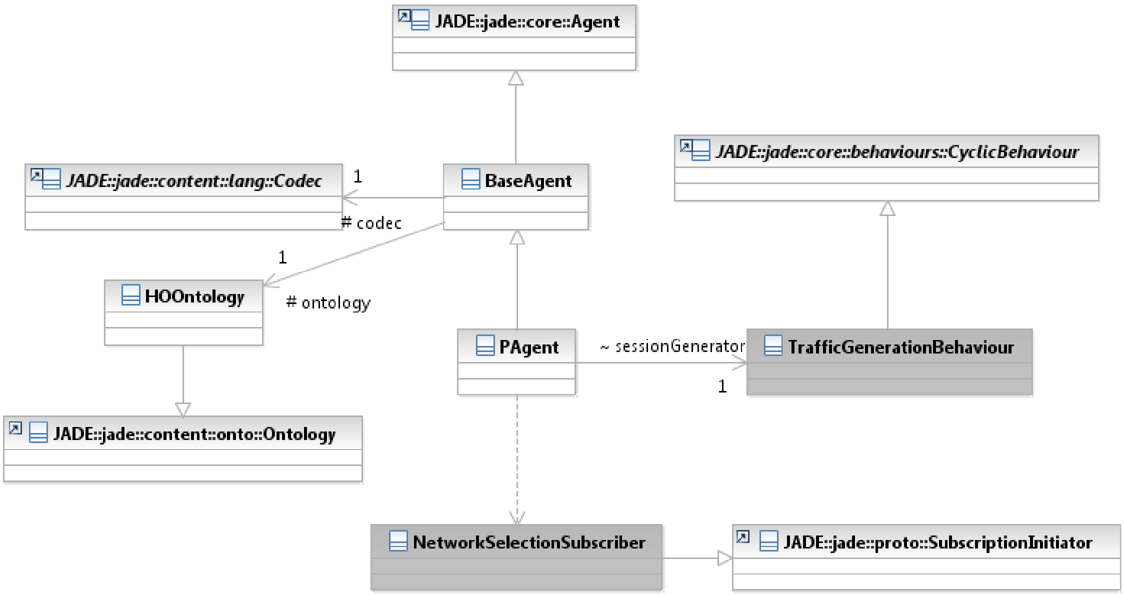
JADE::jade::core::Agent
JADE::jade::content::lang::Codec
1
BaseAgent
# codec
JADE::jade::core::behaviours::CyclicBehaviour
1
HOOntology
# ontology
PAgent
~ sessionGenerator
TrafficGenerationBehaviour
1
JADE::jade::content::onto::Ontology
NetworkSelectionSubscriber
JADE::jade::proto::SubscriptionInitiator

Figure B.1: P-agent design.

P-agent's domain dependent functionality is implemented by `TrafficGenerationBehaviour` and `NetworkSelectionSubscriber` classes that both inherit indirectly from `jade.core.behaviours.Behaviour` class (Figure B.1). Other classes that also inherit from `Behaviour` are presented with dark grey shading in the figures of this section. The execution model of JADE agents is based on behaviours, i.e., non preemptive tasks that are scheduled and sequentially executed while the agent is active in the JADE platform. Each behaviour implementation has an `action` method that includes the task functionality. Once a `Behaviour` instance has been executed, it is either discarded or re-scheduled for execution.

The `TrafficGenerationBehaviour` is executed iteratively and is responsible for simulating the random arrival of VoIP sessions. Each session arrival is announced to AF-agent with an ACL message (*REQ* message of Section 4.5) and the handover decision mechanism is triggered. The `NetworkSelectionSubscriber` behaviour implements the initiator role of the FIPA Subscribe interaction protocol [FIP02e] and is used for subscribing to results of handover decisions performed by the P-agent's corresponding AF-agent. The behaviour is executed after a handover to a new network and the subscription has as target (*participant* role [FIP02e]) the NP-agent. The reason is that handover decision notification messages generated by AF-agents (*RES* messages of Section 4.5) are forwarded to NP-agent that is then responsible for informing P-agents. NP-agent introduces to each message the network's current packet transfer delay that is required for simulation results and is logged by P-agent.

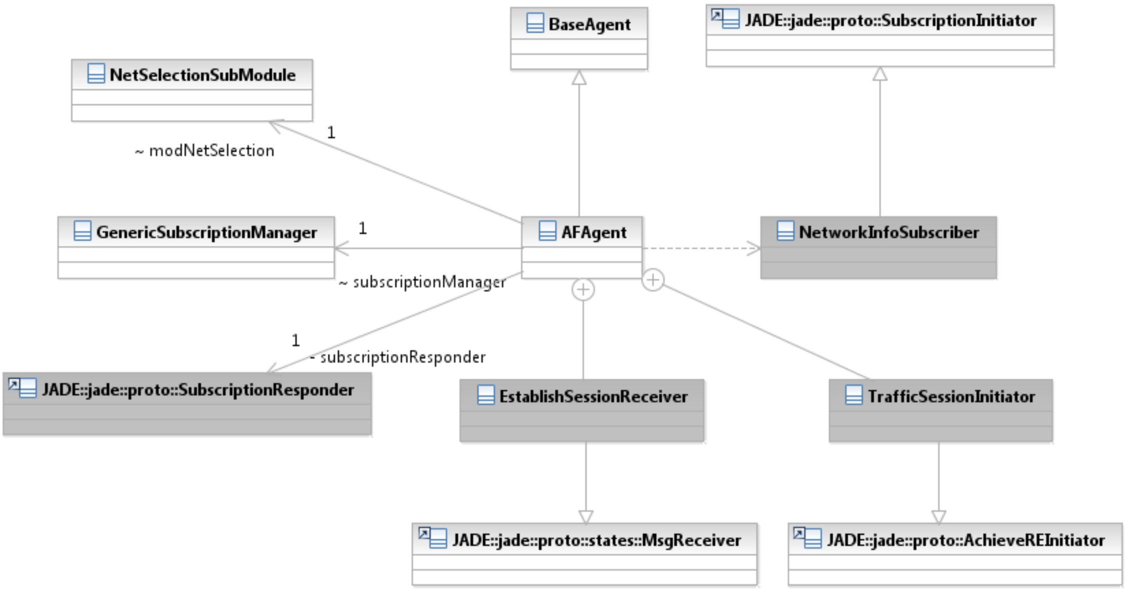
BaseAgent
JADE::jade::proto::SubscriptionInitiator
NetSelectionSubModule
1
~ modNetSelection
GenericSubscriptionManager
1
AFAgent
NetworkInfoSubscriber
~ subscriptionManager
1
- subscriptionResponder
JADE::jade::proto::SubscriptionResponder
EstablishSessionReceiver
TrafficSessionInitiator
JADE::jade::proto::states::MsgReceiver
JADE::jade::proto::AchieveREInitiator

Figure B.2: AF-agent design.

Figure B.2 presents the design of AF-agent. The tasks that are performed by this agent are also Behaviour instances:

- `NetworkInfoSubscriber` that initiates a subscription to NM-agent for network availability and status information and handles the respective notifications. Each notification triggers handover decision in case that a certain threshold of QoS degradation is surpassed. The behaviour implements the initiator role of the FIPA Subscribe interaction protocol.

- `TrafficSessionInitiator` that requests the simulation of a VoIP flow to NP-agent and waits its completion. The behaviour is scheduled upon receiving the *REQ* message from P-agent and implements the *initiator* role of the FIPA Request interaction protocol [FIP02c].

- `EstablishSessionReceiver` that is a permanently active behaviour for handling VoIP session establishment requests from P-agent. The behaviour schedules for execution a `TrafficSessionInitiator` instance upon each request.

- `SubscriptionResponder` that is part of the JADE framework and implements the participant role of the FIPA Subscribe interaction protocol. The behaviour notifies subscribers for network selection events.

The content of each notification sent to subscribers is managed by the `GenericSubscriptionManager` that implements the `SubscriptionManager` interface (Figure B.3). `SubscriptionResponder` opens and closes subscriptions on the basis of messages received by initiators. Moreover, it maintains a list of all active subscriptions represented as instances of `Subscription` class. The agent that executes a `SubscriptionResponder` instance has access to `Subscription` objects and can notify any subscriber by calling the `notify` method of the respective subscription. The `GenericSubscriptionManager` manages the types of subscriptions that can be processed by an agent. The various subscription types correspond to different predicates (queries) that the subscribers are interested to, and are represented as instances of the `SubscriptionModule` interface. Each `SubscriptionModule` instance is kept up to date by the agent with all required information and returns through appropriate methods (a) an ACL message with the current content of the notification, (b) a list of all `Subscription` objects that are relevant to its predicate. These methods are used by the `GenericSubscriptionManager` that is

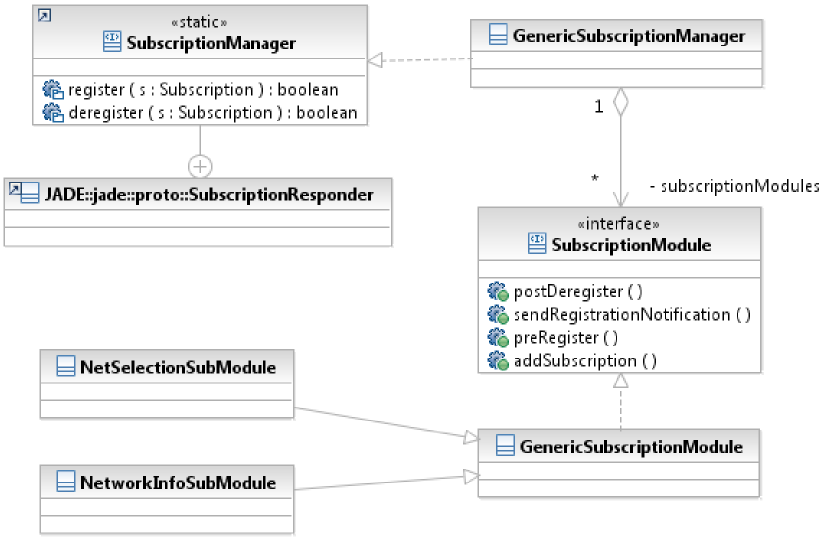
«static»
SubscriptionManager
register (s : Subscription) : boolean
deregister (s : Subscription) : boolean
GenericSubscriptionManager
1
*
- subscriptionModules
JADE::jade::proto::SubscriptionResponder
«interface»
SubscriptionModule
postDeregister ( )
sendRegistrationNotification ( )
preRegister ( )
addSubscription ( )
NetSelectionSubModule
GenericSubscriptionModule
NetworkInfoSubModule

Figure B.3: SubscriptionManager detailed design.

also responsible for adding/removing subscriptions to instances of `SubscriptionModule` through callbacks from `SubscriptionResponder`.

The `GenericSubscriptionManager` exposes a simple interface towards its agent for notifying subscribers. Specifically, the `sendNotifications` method handles the notification of all subscribers that are interested for a certain predicate (subscription type). This predicate is passed as a string argument to the method and determines the `SubscriptionModule` instance that the manager will use for generating notifications. After selecting the appropriate subscription module, the manager receives from it the notification message and the list of active subscriptions and calls the `notify` method of each `Subscription` object. This method uses the `SubscriptionResponder` interface to send an appropriate notification to the corresponding subscriber.

AF-agent uses the `NetSelectionSubModule` subscription module for generating notifications on its handover decisions. Whenever, AF-agent selects a new network it notifies its subscribers on the selected target network. The subscriber for such notifications is the NP-agent of the AF-agent's current platform that forwards this notifications to the corresponding P-agent.

The types of behaviours that are executed by NM-agent are depicted in Figure B.4:

- `NetworkInfoSubscriber` that subscribes to NP-agents for information related to

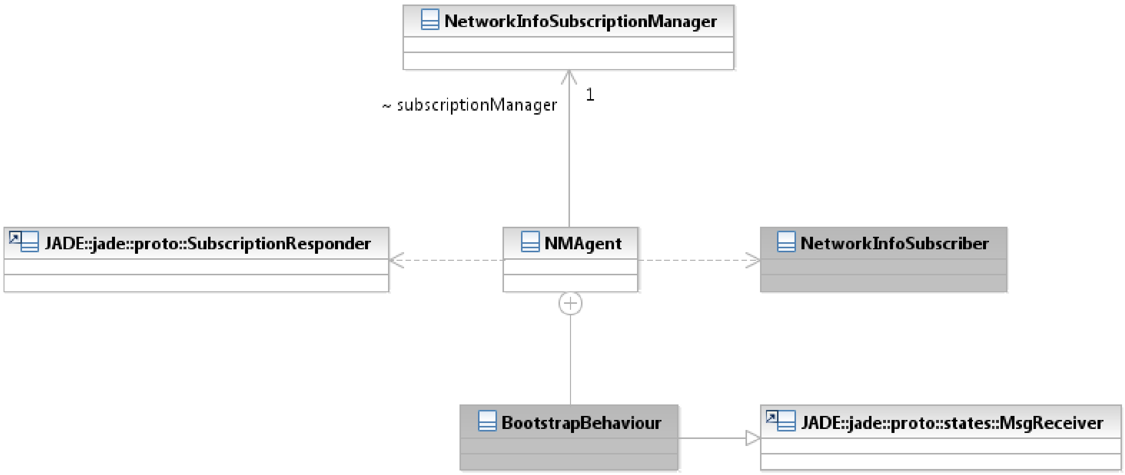
NetworkInfoSubscriptionManager
~ subscriptionManager
1
JADE::jade::proto::SubscriptionResponder
NMAgent
NetworkInfoSubscriber
BootstrapBehaviour
JADE::jade::proto::states::MsgReceiver

Figure B.4: NM-agent design.

the availability and status of their bearer services. The behaviour implements the initiator role of the FIPA Subscribe interaction protocol.

- `SubscriptionResponder` that notifies AF-agent subscribers on the availability and status of networks they are interested to. The behaviour implements the participant role of the FIPA Subscribe interaction protocol.
- `BootStrapBehaviour` that is executed once upon simulation startup and NM-agent instantiation. The behaviour schedules a `NetworkInfoSubscriber` instance for each NP-agent that corresponds to a network that overlaps with the current network's coverage area.

The NP-agent is a key part of the simulation system and incorporates behaviours for interacting with all other agent types. The types of behaviours scheduled in an NP-agent instance are:

- `SimulatorBehaviour` that simulates the duration of VoIP sessions established by P-agents. Moreover, on the basis of active sessions it estimates the network's packet transfer delay.
- `SessionCompleteNotifier` that is scheduled for execution once a session simulated by the `SimulatorBehaviour` has completed. The behaviour informs AF-agent on this fact.

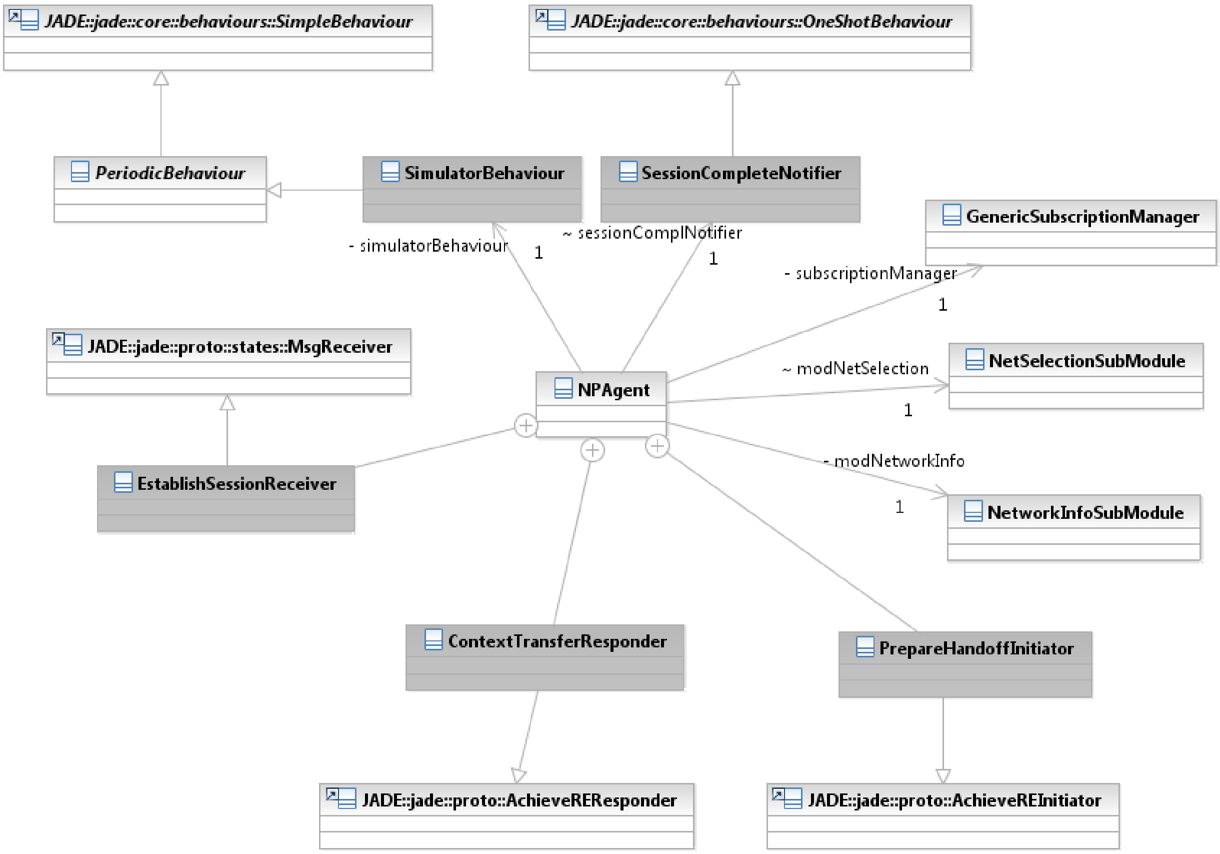

Figure B.5: NP-agent design.

- `EstablishSessionReceiver` that handles session establishment requests by AF-agents and schedules them for simulation.

- `PrepareHandoffInitiator` that is scheduled whenever the NP-agent receives from an AF-agent a network selection notification. The behaviour implements the initiator role of the FIPA Request interaction protocol and initiates a context transfer with the NP-agent that is the target for handover execution. The context information includes the session that will be transferred and simulated by the new NP-agent.

- `ContextTransferResponder` that implements the participant role of the FIPA Request interaction protocol and handles requests generated by the behaviour `PrepareHandoffInitiator`.

The subscription types that are handled by NP-agent are implemented by classes `NetSelectionSubModule` and `NetworkInfoSubModule`. The first generates network selection notifications for P-agents while the latter generates network information notifications for NM-agents.

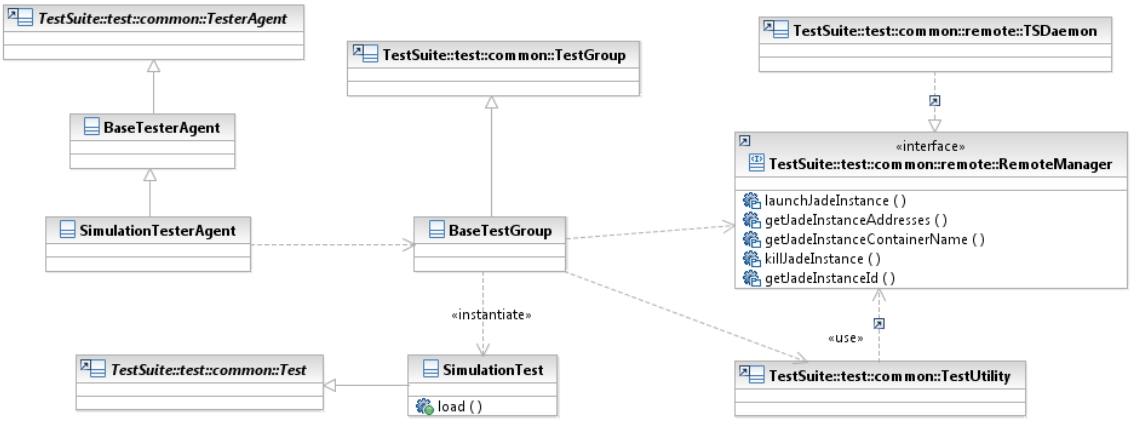
TestSuite::test::common::TesterAgent
BaseTesterAgent
SimulationTesterAgent
TestSuite::test::common::TestGroup
BaseTestGroup
«instantiate»
TestSuite::test::common::Test
SimulationTest
load ()
TestSuite::test::common::remote::TSDaemon
«interface»
TestSuite::test::common::remote::RemoteManager
launchJadeInstance ()
getJadeInstanceAddresses ()
getJadeInstanceContainerName ()
killJadeInstance ()
getJadeInstanceId ()
«use»
TestSuite::test::common::TestUtility

Figure B.6: Infrastructure for launching simulation executions.

The automatic launching of the agent platforms that constitute the simulation system, as well as the instantiation of the various agent types has been implemented on the basis of JADE Test Suite framework [jad10b]. Each simulation scenario has been implemented as a `test.common.Test` instance that is executed by the framework. With regard to the deployment of the simulation system components, the workstation that executes the client platform also executes the test suite framework. The workstations that host the server platforms execute an `rmiregistry` instance with a registered instance of `TSDaemon` that implements the `RemoteManager` interface for remote management of agent platforms.

B.2 Agent mobility overhead evaluation

The agent mobility evaluation simulation has also been implemented as a descendant of `test.common.Test` class and executed through the JADE Test Suite framework. Figure B.7 depicts the design of the agent used during evaluation of performance implications related to agent mobility. The `MobileAgent` has a single behaviour that periodically triggers agent inter-platform migration. The inter-platform migration is supported by the Inter-Platform Mobility Service (IPMS) that is implemented as a JADE add-on [Dep10].

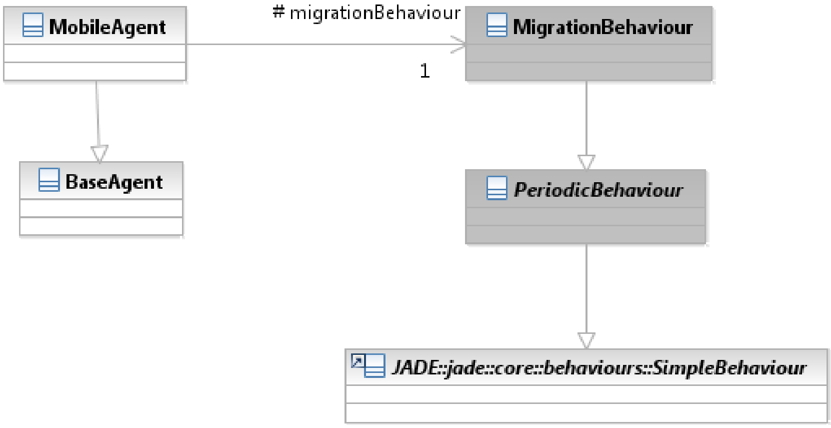

Figure B.7: Software agent used for evaluation of agent mobility overhead.

Bibliography

[3GP04] 3GPP. Feasibility Study for Orthogonal Frequency Division Multiplexing (OFDM) for UTRAN enhancement (Release 6). Technical Report TR 25.892 V6.0.0, 6 2004. 72

[3GP08a] 3GPP. 3GPP system to Wireless Local Area Network (WLAN) interworking;System description (Release 9). Technical Report TS 23.234, 3GPP, 12 2008. 8, 9, 80

[3GP08b] 3GPP. Quality of Service (QoS) concept and architecture (Release 8). Technical Report TS 23.107 V8.0.0, 12 2008. 39

[3GP09a] 3GPP. Architecture enhancements for non-3GPP accesses (Release 9). Technical Report TS 23.402 V9.2.0, 3GPP, 9 2009. 8, 14, 27

[3GP09b] 3GPP. Feasibility study on 3GPP system to Wireless Local Area Network (WLAN) interworking (Release 9). Technical Report TR 22.934, 3GPP, 12 2009. 6

[3GP09c] 3GPP. Generic Access Network (GAN); Stage 2 (Release 9). Technical Report TS 43.318 v9.0.0, 3GPP, 2 2009. 8

[3GP09d] 3GPP. Mobility between 3GPP-Wireless Local Area Network (WLAN) interworking and 3GPP systems (Release 8). Technical Report TS 23.327 V8.4.0 (2009-09), 3GPP, 9 2009. 8

[3GP09e] 3GPP. Multi access PDN connectivity and IP flow mobility. Technical Report TR 23.861 V1.3.0, 3GPP, 5 2009. 14

[3GP09f] 3GPP. Principles of circuit telecommunication services supported by a Public Land Mobile Network (Release 9). Technical Report TS 22.001 v.6.0.0, 3GPP, 12 2009. 89

[3GP10a] 3GPP. Access Network Discovery and Selection Function (ANDSF) Management Object (MO) (Release 9). Technical Report TS 24.312, 3GPP, 3 2010. 27

[3GP10b] 3GPP. Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks; Stage 3; (Release 9). Technical Report TS 24302-920, 3GPP, 3 2010. 9, 27

[3GP10c] 3GPP. IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2 (Release 10). Technical Report TS 23.261 V10.0.0, 3GPP, Jun 2010. 14

[3GP10d] 3GPP. Service aspects; Services and service capabilities (Release 9). Specification TS 22.105 V9.1.0 (2010-09), 3GPP, 2010. 89

[Amb05] Ambient Networks Consortium. Ambient networks contextware-second paper on context-aware networks. Deliverable IST-2002-507134-AN/D6-3, 2005. 35

[AMX05] Ian F. Akyildiz, Shantidev Mohanty, and Jiang Xie. A ubiquitous mobile communication architecture for next-generation heterogeneous wireless systems. *IEEE Communications Magazine*, 43(6):S29–S36, June 2005. DOI: 10.1109/M-COM.2005.1452832. 33

[BCCF05] Paolo Bellavista, Marcello Cinque, Domenico Cotroneo, and Luca Foschini. Integrated support for handoff management and context awareness in heterogeneous wireless networks. In *MPAC '05: Proceedings of the 3rd int. workshop on Middleware for pervasive and ad-hoc computing*, pages 1–8, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-268-2. DOI: 10.1145/1101480.1101495. 31

[BCG09] Paolo Bellavista, Antonio Corradi, and Carlo Giannelli. Mobility-aware Management of Internet Connectivity in Always Best Served Wireless Scenarios. *Springer Mobile Networks and Applications*, 14(1):18–34, February 2009. 2, 20

[BJK$^+$10] CJ. Bernardos, M. Jeyatharan, R. Koodli, T. Melia, and F. Xia. Proxy Mobile IPv6 Extensions to Support Flow Mobility. Intended Standards Track draft-bernardos-netext-pmipv6-flowmob-00, IETF Network-Based Mobility Extensions WG, Jun 2010. 14

[BLH09] Jean-Marie Bonnin, Imed Lassoued, and Zied Ben Hamouda. Automatic multi-interface management through profile handling. *Springer Mobile Networks and Applications*, 14(1):4–17, 2009. 19

[CGG08] Monique Calisti, Roberto Ghizzioli, and Dominic Greenwood. Autonomic service access management for next generation converged networks. In *Advanced Autonomic Networking and Communication*, Whitestein Series in Software Agent Technologies and Autonomic Computing, pages 101–126. Birkhauser Basel, 2008. 34, 35

[Coe00] Carlos A. Coello. An updated survey of GA-based multiobjective optimization techniques. *ACM Comput. Surv.*, 32(2):109–143, 2000. 43

[CR06] K. Chebrolu and R.R. Rao. Bandwidth aggregation for real-time applications in heterogeneous wireless networks. *IEEE Transactions on Mobile Computing*, 5(4):388–403, April 2006. 23, 24

[CVS$^+$04] Rajiv Chakravorty, Pablo Vidales, Kavitha Subramanian, Ian Pratt, and Jon Crowfort. Performance issues with Vertical Handovers - Experiences from GPRS Cellular and WLAN Hot-spots Integration. In *Proceedings of 2nd IEEE Pervasive Computing and Communications Conf.*, pages 155–164, 2004. DOI: 10.1109/PERCOM.2004.1276854. 93, 95

[Dep10] Department of Information and Communications Engineering, Autonomous University of Barcelona (UAB) . JADE Inter-Platform Mobility Project. https://tao.uab.cat/ipmp/, 2010. 95, 125

[DKK$^+$00] M. Dawande, J. Kalagnanam, P. Keskinocak, F.S. Salman, and R. Ravi. Approximation Algorithms for the Multiple Knapsack Problem with Assignment Restrictions. *Journal of Combinatorial Optimization*, 4:171–186, 2000. 10.1023/A:1009894503716. 22, 38, 48, 49

[DMT10] DMTF Networks Working Group. Network Specification. CIM Schema Specification v. 2.25.0, 3 2010. 90

[EMWK08] T. Ernst, N. Montavont, R. Wakikawa, and K. Kuladinithi. Motivations and Scenarios for Using Multiple Interfaces and Global Addresses. Internet-Draft, IETF MONAMI6 Working Group, May 2008. 2, 10, 36

[FCCM07] Roberta Fracchia, Claudio Casetti, Carla-Fabiana Chiasserini, and Michela Meo. WiSE: Best-path selection in Wireless Multihoming Environments. *IEEE Transactions on Mobile Computing*, 6(10):1130–1141, oct 2007. 23, 24

[FIP02a] FIPA. FIPA ACL Message Structure Specification. Standard specification, Foundation for Intelligent Physical Agents, 2002. 80

[FIP02b] FIPA. FIPA Iterated Contract Net Interaction Protocol Specification. Standard specification, Foundation for Intelligent Physical Agents, 2002. 35

[FIP02c] FIPA. FIPA Request Interaction Protocol Specification. Standard specification, Foundation for Intelligent Physical Agents, 2002. 121

[FIP02d] FIPA. FIPA SL Content Language Specification. Standard specification, 2002. http://www.fipa.org/specs/fipa00008/SC00008I.html. 119

[FIP02e] FIPA. FIPA Subscribe Interaction Protocol Specification. Standard specification, Foundation for Intelligent Physical Agents, 2002. 120

[FIP10] FIPA. FIPA Specifications. Specifications, 2010. www.fipa.org/specifications/index.html. 80, 82, 118

[GAM05] V Gazis, N Alonistioti, and L Merakos. Toward a Generic "Always Best Connected" Capability in Integrated WLAN/UMTS Cellular Mobile Networks (and Beyond). *IEEE Wireless Communications Magazine*, 12(3):20–29, 2005. 21, 22

[GJ90] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990. ISBN 0716710455. 49, 50

[GJ03] Eva Gustafsson and Annika Jonsson. Always Best Connected. *IEEE Wireless Communications Magazine*, 10(1):49–55, February 2003. DOI: 10.1109/MWC.2003.1182111. 2, 5, 78, 79

[GLD$^+$08] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil. Proxy Mobile IPv6. IETF Standards Track RFC 5213, IETF Network Working Group, 8 2008. 12

[GQX$^+$04] David K. Goldenberg, Lili Qiuy, Haiyong Xie, Yang Richard Yang, and Yin Zhang. Optimizing cost and performance for multihoming. *SIGCOMM Comput. Commun. Rev.*, 34(4):79–92, 2004. 18

[HNH07] Ahmed Hasswa, Nidal Nasser, and Hossam Hassanein. A seamless context-aware architecture for fourth generation wireless networks. *Wirel. Pers. Commun.*, 43: 1035–1049, November 2007. 32

[IAS06] J.R. Iyengar, P.D. Amer, and R. Stewart. Concurrent Multipath Transfer Using SCTP Multihoming Over Independent End-to-End Paths. *IEEE/ACM Transactions on Networking*, 14(5):951–964, Oct. 2006. 23, 24

[IEE07] IEEE 802.21 Working Group. Media Independent Handover Services. Draft standard, IEEE, 2007. 39

[IEE08] IEEE. Ieee standard for local and metropolitan area networks - part 21: Media independent handover services. Technical Report Std 802.21-2008, IEEE, 2008. 25

[IEE09a] IEEE. IEEE Standard for architectural building blocks enabling network-device distributed decision making for optimized radio resource usage in heterogeneous wireless access networks. *IEEE Std 1900.4-2009*, pages C1–119, 2009. 1, 28, 29

[IEE09b] IEEE TGu. Draft Amendment to Standard for Information Technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 7: Interworking with External Networks. *IEEE Unapproved Draft Std P802.11u/D8.0, July 2009*, 2009. 31

[IEE10a] IEEE. Standard for Local and Metropolitan Area Networks: IEEE Media Independent Handover Services - Amendment: Optimized Single Radio Handovers. Technical report, 2010. 7

[IEE10b] IEEE. Status of Project IEEE 802.11u Interworking with External Networks. Technical report, 2010. 9

[jad10a] JADE - Java Agent DEvelopment Framework. http://jade.tilab.com, 2010. 93, 118

[jad10b] JADE Test Suite user guide. http://jade.tilab.com/doc/tutorials/JADE_TestSuite.pdf, 2010. 125

[KAE07] Nihat Kasap, Haldun Aytug, and S. Selcuk Erenguc. Provider selection and task allocation issues in networks with different QoS levels and all you can send pricing. *Elsevier Decision Support Systems*, 43(2):375–389, 2007. 21, 22

[KH03] Geir M Koien and Thomas Haslestad. Security Aspects of 3G-WLAN Interworking. *IEEE Communications Magazine*, 41(11):82–88, November 2003. DOI: 10.1109/MCOM.2003.1244927. 92

[Kle75] Leonard Kleinrock. *Theory, Volume 1, Queueing Systems.* Wiley-Interscience, 1975. ISBN 0471491101. 95

[KS07] Kyu-Han Kim and Kang G. Shin. PRISM: Improving the Performance of Inverse-Multiplexed TCP in Wireless Networks. *IEEE Transactions on Mobile Computing*, 6:1297–1312, 2007. 23, 24

[KZSH05] Kyu-Han Kim, Yujie Zhu, Raghupathy Sivakumar, and Hung-Yun Hsieh. A receiver-centric transport protocol for mobile hosts with heterogeneous wireless interfaces. *Wirel. Netw.*, 11(4):363–382, 2005. 23, 24

[LFP99] Y Labrou, T Finin, and Y Peng. Agent Communication Languages: The current landscape. *IEEE Intelligent Systems*, 14(2):45–52, March/April 1999. DOI: 10.1109/5254.757631. 35, 80, 82

[LIN09] LINDO Systems. LINGO - Optimization Modeling Software for Linear, Nonlinear, and Integer Programming. http://www.lindo.com, 2009. 15, 62

[LN03] J. Lorchat and T. Noel. Power performance comparison of heterogeneous wireless network interfaces. In *IEEE 58th Veh. Technology Conf., (VTC'03-Fall)*, pages 2182–2186, 2003. 39, 40

[LWZ08] Jianxin Liao, Jingyu Wang, and Xiaomin Zhu. A multi-path mechanism for reliable VoIP transmission over wireless networks. *Elsevier Computer Networks*, 52(13):2450–2460, September 2008. 23, 24

[LX07] Xiaomei Liu and Li Xiao. A Survey of Multihoming Technology in Stub Networks: Current Research and Open Issues. *IEEE Network*, 21(3):32–40, 2007. 17

[MG10] T. Melia and S. Gundavelli. Logical Interface Support for multi-mode IP Hosts. Internet Draft (Informational) draft-ietf-netext-logical-interface-support-00, IETF Network-Based Mobility Extensions WG, Aug 2010. 14

[MT90] Silvano Martello and Paolo Toth. *Knapsack Problems: Algorithms and Computer Implementations.* John Wiley and Sons, 1990. 48

[Mus09] Mushroom Networks Inc. Wireless broadband bonding network appliance. http://www.mushroomnetworks.com/, 2009. 3, 36

[MWE$^+$08] N. Montavont, R. Wakikawa, T. Ernst, C. Ng, and K. Kuladinithi. Analysis of Multihoming in Mobile IPv6. Internet-draft(informational), IETF, May 2008. 10

[NVAGD08] Quoc-Thinh Nguyen-Vuong, Nazim Agoulmine, and Yacine Ghamri-Doudane. A user-centric and context-aware solution to interface management and access network selection in heterogeneous wireless environments. *Elsevier Computer Networks*, 52:3358•–3372, Sep 2008. 19, 72, 73, 74, 108

[OMA08a] OMA. Standardized Connectivity Management Objects. Technical Report OMA-DDS-DM_ConnMO-V1_0- 20081107-A, Open Mobile Alliance, 2008. 28

[OMA08b] OMA. Standardized Connectivity Management Objects WLAN Parameters. Technical Report OMA-DDS-DM_ConnMO_WLAN-V1_0-20081024-A, Open Mobile Alliance, 2008. 28

[RN03] Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach, Second Edition*, pages 94–136. Prentice Hall, 2003. 45, 52

[RT07] M. Riegel and M. Tuexen. Mobile sctp. Internet draft (experimental), IETF Network Working Group, Nov. 2007. 32

[SJ05] Qingyang Song and Abbas Jamalipour. An adaptive quality-of-service network selection mechanism for heterogeneous mobile networks. *Wireless Communications and Mobile Computing*, 5(6):697–708, 2005. DOI: 10.1002/wcm.330. 91, 94

[SK98] Mark Stemm and Randy H. Katz. Vertical handoffs in wireless overlay networks. *Mobile Networks and Applications*, 3(4):335–350, 1998. 6

[SKK08] Anne Setamaa-Karkkainen and Jani Kurhinen. Optimal usage of multiple network connections. In *MOBILWARE '08: Proceedings of the 1st international conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications*, pages 1–6, 2008. 23

[SNW06] E. Stevens-Navarro and V.W.S. Wong. Comparison between vertical handoff decision algorithms for heterogeneous wireless networks. *IEEE 63rd Vehicular Technology Conference (VTC'06)*, 2:947–951, May 2006. 73

[Sol09] H. Soliman. Mobile IPv6 Support for Dual Stack Hosts and Routers. IETF Standards Track RFC 5555, IETF Network Working Group, Jun 2009. 14

[SRJS05] Jun Zhao Sun, Jukka Riekki, Marko Jurmu, and Jaakko Sauvola. Adaptive connectivity management middleware for heterogeneous wireless networks. *IEEE Wireless Communications Magazine*, 12(6):18–25, December 2005. DOI: 10.1109/MWC.2005.1561941. 32

[STM$^+$10] H. Soliman, G. Tsirtsis, N. Montavont, G. Giaretta, and K. Kuladinithi. Flow Bindings in Mobile IPv6 and Nemo Basic Support. Internet Draft draft-ietf-mext-flow-binding-06.txt, IETF MEXT Working Group, 3 2010. 12, 14, 37

[SXM$^+$00] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson. Stream Control Transmission Protocol. Standard RFC 2960, IETF Network Working Group, oct 2000. 24

[SZ06] Farhan Siddiqui and Sherali Zeadally. Mobility management across hybrid wireless networks: Trends and challenges. *Computer Communications*, 29(9):1363–1385, August 2006. DOI: 10.1016/j.comcom.2005.09.003. 77

[TGSM10] G. Tsirtsis, G. Giarreta, H. Soliman, and N. Montavont. Traffic Selectors for Flow Bindings. Internet Draft draft-ietf-mext-binary-ts-04.txt, IETF Network Working Group, 2 2010. 12

[THL06] N. Thompson, G. He, and H. Luo. Flow Scheduling for End-Host Multihoming. In *Proc. IEEE Int. Conf. on Computer Communications (INFOCOM'06)*, pages 1–12, apr 2006. 18

[TS09] Cheng-Lin Tsao and Raghupathy Sivakumar. On effectively exploiting multiple wireless interfaces in mobile hosts. In *CoNEXT '09: Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 337–348, New York, NY, USA, 2009. ACM. 23, 24

[VBS+05] Pablo Vidales, Javier Baliosian, Joan Serrat, Glenford Mapp, Frank Stajano, and Andy Hopper. Autonomic system for mobility support in 4G networks. *IEEE Journal on Selected Areas on Communications*, 23(12):2288– 2304, December 2005. DOI: 10.1109/JSAC.2005.857198. 33

[VZ10] Dimitris El. Vassis and Vassilis E. Zafeiris. Pamvotis - IEEE 802.11 WLAN Simulator. URL http://www.pamvotis.org, 2010. 94

[WDT+09] R. Wakikawa, V. Devarapalli, G. Tsirtsis, T. Ernst, and K. Nagami. Multiple Care-of Addresses Registration. Old Internet Draft (Proposed Standard) RFC 5648, IETF Network Working Group, 10 2009. 12, 14, 37

[WFP+06] Qing Wei, Karoly Farkas, Christian Prehofer, Paulo Mendes, and Bernhard Plattner. Context-aware handover using active network technology. *Computer Networks*, 50(15):2855–2872, October 2006. DOI: 10.1016/j.comnet.2005.11.002. 34

[WiM08] WiMAX Forum. WiMAX - 3GPP Interworking. Technical Report T37-002-R010v3, WiMAX Forum, 1 2008. 9

[WWY+07] Shupeng Wang, Jianping Wang, Mei Yang, Xiaochun Yun, and Yingtao Jiang. Handover cost optimization in traffic management for multi-homed mobile networks. *Ubiquitous Intelligence and Computing*, pages 295–308, 2007. 22

[XV05] B Xing and Nalini Venkatasubramanian. Multi-constraint dynamic access selection in always best connected networks. In *Proc. Int. Conf. on Mobile and Ubiquitous Systems (Mobiquitous'05)*, pages 56–64, 2005. 21, 22, 39, 40, 91

[Zah03] T Zahariadis. Trends in the path to 4G. *IEE Communications Engineer Magazine*, 1(1):12–15, February 2003. 77

[ZG08] Vassilis E. Zafeiris and E. A. Giakoumakis. Towards flow scheduling optimization in multihomed mobile hosts. In *Proc. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC 2008)*, pages 1–5, Sept. 2008. 45, 67