NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

PhD THESIS

Management of Network and Energy Resources in Cognitive and Self-Organizing Wireless Networks

Apostolos D. Kousaridas

ATHENS

FEBRUARY 2013

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Διαχείριση Δικτυακών και Ενεργειακών Πόρων σε Γνωσιακά και Αυτο-οργανωμένα Ασύρματα Δίκτυα

Απόστολος Δ. Κουσαρίδας

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2013

PhD THESIS

Management of Network and Energy Resources in Cognitive and Self-Organizing Wireless Networks

Apostolos D. Kousaridas

SUPERVISOR: Lazaros Merakos, NKUA Professor

THREE-MEMBER ADVISORY COMMITTEE:
Lazaros Merakos, NKUA Professor
Theodoros Apostolopoulos, AUEB Professor
Efstathios Hadjiefthymiades, NKUA Associate Professor

SEVEN-MEMBER EXAMINATION COMMITTEE

Lazaros Merakos,
NKUA Professor

Theodoros Apostolopoulos,
AUEB Professor

Efstathios Hadjiefthymiades,
NKUA Associate Professor

Ioannis Stavrakakis,
NKUA Professor

Manolis Koubarakis,
NKUA Professor

Athanasia (Nancy) Alonistioti,
NKUA Lecturer

Michael Theologou,
NTUA Professor

Examination Date 20/02/2013

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Διαχείριση Δικτυακών και Ενεργειακών Πόρων σε Γνωσιακά και Αυτο-οργανωμένα Ασύρματα Δίκτυα

Απόστολος Δ. Κουσαρίδας

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Λάζαρος Μεράκος, Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:
Λάζαρος Μεράκος, Καθηγητής ΕΚΠΑ
Θεόδωρος Αποστολόπουλος, Καθηγητής ΟΠΑ
Ευστάθιος Χατζηευθυμιάδης, Αναπληρωτής Καθηγητής ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Λάζαρος Μεράκος,
Καθηγητής ΕΚΠΑ

Θεόδωρος Αποστολόπουλος,
Καθηγητής ΟΠΑ

Ευστάθιος Χατζηευθυμιάδης,
Αναπληρωτής Καθηγητής ΕΚΠΑ

Ιωάννης Σταυρακάκης,
Καθηγητής ΕΚΠΑ

Μανόλης Κουμπαράκης,
Καθηγητής ΕΚΠΑ

Αθανασία Αλωνιστιώτη,
Λέκτορας ΕΚΠΑ

Μιχαήλ Θεολόγου,
Καθηγητής ΕΜΠ

Ημερομηνία εξέτασης 20/02/2013

ABSTRACT

The efficient usage of network resources and the reduction of the consumed energy in modern communication systems that operate in a dense urban environment are challenging tasks, due to the complexity and the spatio-temporal dynamics of wireless networks. Self-organization is considered as one of the most promising paradigms for the management of networks that operate in highly dynamic and dense environments. In this thesis a novel approach is proposed to dynamically control the size and configuration of a wireless network for the effective utilization of network resources and energy saving. Firstly, we define the functional and software architecture for a self-organizing network, based on the cognitive cycle concept. Central to our approach is the organization of access points into clusters to facilitate local management and coordination of the deployed cognitive managers. In each cluster, the elected cluster head monitors usage of resources as well as the energy consumption changes during the transmission and reception, at both the access point and user equipment sides, and decides on the appropriate adaptation action. Specifically, a scheme for access point dynamic deactivation and reactivation is proposed for efficient use of radio resources and reduction of access point energy consumption. Also, re-allocation of selected frequency channels and formation of multi-hop relays among user equipment is used to diminish the effect of access point deactivation on user equipment energy consumption. Configuration and performance optimization tasks affect energy consumption of specific components and energy-related metrics of different devices. The energy consumption improvement attained under the proposed scheme, at both the access point and the user equipment sides, is evaluated via simulation. In addition a scheme has been proposed and evaluated for the coordination of different self-optimization actions, which are concurrently triggered by the cognitive managers in a self-organizing network (coverage and capacity optimization, energy saving, interference). The goal is to resolve conflicts on configuration actions and dependencies among performance metrics, ensuring network stability. Finally, an algorithmic framework is proposed, in the context of a single network operator environment, to enhance the effectiveness of cognitive managers. Specifically, the accumulated history of previous events and a feedback-based scheme are introduced for the evolution of the decision making algorithms that are used for the improvement of VoIP QoS in a congested WiMAX network. The proposed algorithmic framework has been evaluated using the FIRE Panlab experimental facilities.

SUBJECT AREA: Communication Networks

KEYWORDS: π.χ. network management, wireless networks, resource management, energy saving, self organization, cognition.

ΠΕΡΙΛΗΨΗ

Η αποδοτική χρήση των δικτυακών πόρων και η ελάττωση της καταναλισκόμενης ενέργειας των συστημάτων επικοινωνιών τα οποία βρίσκονται σε ένα πυκνό αστικό περιβάλλον αποτελούν σημαντικές προκλήσεις, εξαιτίας της πολυπλοκότητας και της χωρο-χρονικής δυναμικής των ασύρματων δικτύων. Η αυτό-οργάνωση των δικτυακών συστημάτων θεωρείται ως μια πολλά υποσχόμενη τεχνική για τη διαχείριση ενός πυκνού ασύρματου δικτύου που λειτουργεί σε πολύπλοκες και συνεχώς μεταβαλλόμενες συνθήκες. Στην παρούσα διατριβή προτείνεται μια καινοτόμος λύση για το δυναμικό έλεγχο του μεγέθους αλλά και της διαμόρφωσης ενός ασύρματου τοπικού δικτύου, με σκοπό την αποτελεσματική χρησιμοποίηση των δικτυακών πόρων και την εξοικονόμηση ενέργειας. Αρχικά, παρουσιάζεται η αρχιτεκτονική ενός αυτό-οργανώμενου δικτύου, η οποία βασίζεται στην ύπαρξη του γνωσιακού κύκλου. Κεντρική ιδέα της διατριβής είναι η οργάνωση των ασύρματων σταθμών βάσης σε συστάδες, οι οποίες διευκολύνουν το συντονισμό και τη τοπική διαχείριση των γνωσιακών διαχειριστών που είναι εγκατεστημένοι στα ασύρματα σημεία πρόσβασης. Σε κάθε σχηματιζόμενη συστάδα, ο επικεφαλής παρακολουθεί το επίπεδο χρησιμοποίησης των διαθέσιμων πόρων καθώς επίσης και τις αλλαγές στην καταναλισκόμενη ενέργεια κατά τη μετάδοση και τη λήψη δεδομένων, τόσο στην πλευρά των σταθμών βάσης όσο και σε αυτή των συσκευών των χρηστών. Αξιολογώντας τις παραπάνω πληροφορίες, επιλέγεται η κατάλληλη ενέργεια αναδιαμόρφωσης του δικτύου. Συγκεκριμένα, προτείνεται ένα σχήμα για τη δυναμική απενεργοποίηση ή ενεργοποίηση σταθμών βάσης με στόχο την εξοικονόμηση ενέργειας στην πλευρά του δικτύου και τη λελογισμένη χρήση των ραδιο-πόρων. Επιπλέον, αλγόριθμοι για την ανακατανομή των επιλεγμένων καναλιών/συχνοτήτων, σύμφωνα με το φόρτο κάθε σταθμού βάσης καθώς και το σχηματισμό μονοπατιών που αποτελούνται από πολλά άλματα, έχοντας ως ενδιάμεσους κόμβους (ή αναμεταδότες) τερματικές συσκευές (συσκευές χρηστών), προτείνονται και χρησιμοποιούνται. Στόχος είναι να περιορισθεί η επίδραση της απενεργοποίησης ενός σταθμού βάσης, στην κατανάλωση της συσκευής του χρήστη. Οι δράσεις για την αυτο-βελτίωση ή την αυτο-διαμόρφωση ενός δικτυακού κόμβου επηρεάζουν την καταναλισκόμενη ενέργεια ενός συγκεκριμένου στοιχείου αλλά παράλληλα επιδρούν και στις μετρικές ενέργειας διαφορετικών συσκευών. Η εξοικονόμηση που επιτυγχάνεται στην καταναλισκόμενη ενέργεια με τη χρήση του προτεινόμενου σχήματος, τόσο στην πλευρά των σταθμών βάσης όσο και των συσκευών του χρήστη, αξιολογείται μέσω πειραμάτων προσομοίωσης. Στη συνέχεια προτείνεται και αποτιμάται ένα αλγοριθμικό σχήμα για το συντονισμό των διαφόρων αναδιαμορφώσεων που προκύπτουν ταυτόχρονα σε έναν ή περισσότερους γνωσιακούς διαχειριστές ενός αυτο-οργανωμένου δικτύου, από διαφορετικά προβλήματα αυτο-βελτίωσης (π.χ., εξορθολογισμός της χωρητικότητας και της παρεχόμενης κάλυψης, ελάττωση διακαναλικών παρεμβολών, εξοικονόμηση ενέργειας). Σκοπός είναι ο εντοπισμός και η επίλυση συγκρούσεων στις αναδιαμορφώσεις που ενεργοποιούνται καθώς και η συνολικά καλύτερη απόδοση του δικτύου, επιζητώντας τη μικρότερη δυνατή αρνητική επίδραση στις υπόλοιπες μετρικές απόδοσης. Τέλος, στα πλαίσια ενός αυτό-διαχειριζόμενου δικτύου, το οποίο ανήκει σε έναν πάροχο, εισάγεται ένα αλγοριθμικό πλαίσιο για την ενίσχυση της αποτελεσματικότητας των γνωσιακών διαχειριστών. Συγκεκριμένα, το ιστορικό προηγούμενων ενεργειών αλλά και ο προτεινόμενος μηχανισμός μάθησης βασιζόμενος στην ανατροφοδότηση εξελίσσουν τον αλγόριθμο λήψης αποφάσεων. Το εν λόγω πλαίσιο χρησιμοποιήθηκε για τη βελτίωση της ποιότητας της υπηρεσία μεταφοράς φωνής μέσω διαδικτύου σε ένα ασύρματο δίκτυο ευρείας περιοχής, ενώ η αποτίμηση της απόδοσής του έγινε στην πειραματική πλατφόρμα FIRE Panlab.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα Επικοινωνιών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: π.χ. διαχείριση δικτύων, ασύρματα δίκτυα, διαχείριση πόρων, εξοικονόμηση ενέργειας, αυτό-οργάνωση, γνωσιακή επιστήμη

I would like to dedicate this dissertation to my parents Dimitris and Aglaia, as well as to Maria and Eleni.

"Vitality brief, art long, occasion precipitous, experience perilous, judgment difficult", Hippocrates (460-377 B.C.)

Στους γονείς μου Δημήτρη και Αγλαΐα, στην Μαρία και στην Ελένη.

«Ο βίος βραχύς, η δε τέχνη μακρή, ο δε καιρός οξύς, η δε πείρα σφαλερή, η δε κρίσις χαλεπή», Ιπποκράτης (460-377 π.Χ.)

.

ACKNOWLEDGEMENTS

At the outset, I would like to express my sincere gratitude to my thesis advisory committee: Professor Lazaros Merakos, Professor Theodoros Apostolopoulos and Associate Professor Efstathios Hadjiefthymiades.

I would like to thank my advisor, Professor Lazaros Merakos for guiding and supporting me. Our academic interactions have always been very beneficial and thought provoking. You have set an example of excellence as a researcher, instructor and mentor.

I would like to express my appreciation to my co-advisor Lecturer Nancy Alonistioti for the excellent collaboration, the guidance through this process and the multi-facet pan-European experience. Your support, ideas, and feedback have been absolutely invaluable for this thesis.

Moreover, I would like to thank Professor Theodoros Apostolopoulos for his sincere help and guidance, even from my undergraduate studies. I will always be grateful for your assistance, which encouraged me to get involved with research.

I would like to sincerely thank my friends and (ex or present) colleagues Dr. Panagis Magdalinos, Dr. Eleni Patouni, Panagiotis Spapis, Dr. Zacharias Boufidis, Dr. Andreas Merentitis, Assistant Prof. Alexandros Kaloxylos, Antonis Lilis, Dimitris Makris, Thanos Makris, Christos Papazafeiropoulos, Evangelos Rekkas, Dr. Spyros Panagiotakis, Dr. Maria Koutsopoulou, Makis Stamatelatos, Costas Polychronopoulos, Aristotelis Glentis. Furthermore, I would like to thank the younger members of the SCAN laboratory George Katsikas, Costas Chatzikokolakis, Roi Arapoglou, Stefanos Falangitis, and Dimitris Charalampous. You have all directly or indirectly contributed in the shaping of this thesis. I will never forget the endless and constructive debates we had for technological, philosophical, societal and historical issues, as well as the great times we had during breaks, recreational activities or even in islands.

I want to thank all members of SCAN laboratory whose friendly company, during my thesis work, was a source of great pleasure. I also thank the partners form the various European projects that I have been involved during my thesis work (E2R, E3, Self-NET, CONSERN, UniverSelf) and especially Dr. Andrej Mihailovic, Dr. Gerard Nguengang, and Dr. Ioannis Chochliouros. I want to dedicate this dissertation to the memory of George, Pantelis, and Prof. Efthimiadis.

Last but not least, I want to thank my parents Dimitris and Aglaia, my sister Maria and Eleni for their love, help and support all these years.

ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Η παρούσα διδακτορική διατριβή ασχολείται με τη θεματική περιοχή των αυτο-οργανωμένων και γνωσιακών συστημάτων επικοινωνιών για τη βελτίωση της διαχείρισης των ραδιο-πόρων και την ελάττωση της καταναλισκόμενης ενέργειας σε περιβάλλοντα ασύρματων επικοινωνιών. Συγκεκριμένα, προδιαγράφει και σχεδιάζει τους απαραίτητους υποστηρικτικούς μηχανισμούς για την ανάπτυξη ενός αυτό-οργανωμένου δικτύου, και ακολούθως προτείνει και αποτιμά ένα σύνολο από αλγορίθμους για την αναδιαμόρφωση παραμέτρων του δικτύου (απενεργοποίηση ή ενεργοποίηση σταθμού βάσης και μεταπομπή χρηστών, αλλαγή επιλεγμένης συχνότητας/καναλιού, σχηματισμός μονοπατιών πολλών αλμάτων μεταξύ των τερματικών σταθμών). Η λήψη απόφασης για την αναδιαμόρφωσης συσχετίζει περιοδικά μετρικές απόδοσης του ασύρματου δικτύου με τους κανόνες πολιτικής και τα κατώφλια που έχει καθορίσει ο διαχειριστής ή ο κατασκευαστής του δικτύου. Στόχος των μηχανισμών αυτο-οργάνωσης είναι αφενός η ελάττωση της παρεμβατικότητας του διαχειριστή του δικτύου, λόγω της αυξανόμενης πολυπλοκότητας διασφαλίζοντας, όμως, την ορθή και αξιόπιστη διαμόρφωσή του, και αφετέρου η βελτίωση και ο εξορθολογισμός της χωρητικότητας του συστήματος, της παρεχόμενης κάλυψης, των διακαναλικών παρεμβολών όσον αφορά στο ρυθμό απώλειας πακέτων δεδομένων, καθώς και της κατανάλωσης ενέργειας.

Η σχεδίαση ενός αυτο-οργανωμένου δικτύου περιλαμβάνει τόσο την αρχιτεκτονική του δομή όσο και τη μορφή της οργάνωσής του. Στη διατριβή περιγράφεται το αρχιτεκτονικό πλαίσιο των οντοτήτων που υλοποιούν ένα αυτο-οργανώμενο δίκτυο. Συγκεκριμένα αναλύονται οι λειτουργίες και οι ικανότητες του γνωσιακού διαχειριστή δικτυακών στοιχείων (επίπεδο Ι - π.χ. σταθμών βάσης) καθώς και του γνωσιακού διαχειριστή για κάθε δικτυακό τομέα (επίπεδο ΙΙ), ο οποίος επιβλέπει ένα σύνολο από διαχειριστές δικτυακών στοιχείων. Όσον αφορά την οργανωτική δομή των παραπάνω γνωσιακών διαχειριστών, στη διατριβή προτείνεται ο αλγόριθμος για τη δημιουργία ενός (υβριδικού) σχήματος συστάδων μεταξύ των κόμβων του ασύρματου τοπικού δικτύου (γνωσιακούς διαχειριστές επιπέδου Ι). Ο συγκεκριμένος αλγόριθμος βασίζεται στα χαρακτηριστικά του σχηματιζόμενου γράφου της περιοχής του δικτύου, και στον κανόνα της “κατά προτίμησης προσκόλλησης”, για να εκλέξει τους επικεφαλείς των ομάδων και να καθορίσει τα σύνορα της κάθε συστάδας. Οι εκλεγμένοι αρχηγοί αποτελούν τους γνωσιακούς διαχειριστές τομέα (επίπεδο ΙΙ). Σκοπός των συστάδων είναι η εγκαθίδρυση συνεργατικών αυτο-διαχειριζόμενων δομών οι οποίες διευκολύνουν την επίλυση προβλημάτων διαχείρισης πόρων σε μια ευρύτερη περιοχή του δικτύου και την αποφυγή συγκρούσεων μεταξύ των κόμβων όσο αφορά τις επιμέρους αυτο-βελτιώσεις. Σε αυτό το σημείο αξίζει να σημειωθεί ότι η εισαγωγή ικανοτήτων αυτοδιαχείρισης αυξάνει την αυτονομία και την αποδοτικότητα, αλλά αναπόφευκτα αυξάνει τους αναγκαίους πόρους σε CPU και RAM σε κάθε κόμβο όπως και την κίνηση στο τοπικό δίκτυο από τις ανταλλαγές δεδομένων παρακολούθησης κλπ.

Αξιοποιώντας τις ιδιότητες και τα χαρακτηριστικά της αυτο-οργάνωσης προτείνεται ένα σύνολο από αλγορίθμους για τη διαχείριση των πόρων των ασύρματων δικτύων. Αρχικά παρουσιάζεται ένας καινοτόμος αλγόριθμος για τη δυναμική ενεργοποίηση και απενεργοποίηση σταθμών βάσης και τη μεταπομπή των συνδεδεμένων τερματικών, που αποσκοπεί στην εξορθολογισμένη χρήση των ραδιο-πόρων βάσει της πυκνότητας και της κίνησης του δικτύου. Επιπλέον, προτείνεται ένας αλγόριθμος για το σχηματισμό σύντομων διαδρομών/μονοπατιών για την αποστολή πακέτων δεδομένων που αποτελούνται από πολλά άλματα, έχοντας ως ενδιάμεσους κόμβους (ή αναμεταδότες) τερματικές συσκευές (συσκευές χρηστών). Επίσης, παρουσιάζεται ένα σχήμα επιλογής καναλιών/συχνοτήτων σύμφωνα με το φόρτο κάθε σταθμού βάσης. Τα αποτελέσματα της προσομοίωσης σε OPNET δείχνουν το κέρδος τόσο στους πόρους που εξοικονομούνται όσο και στην καταναλισκόμενη ενέργεια από την απενεργοποίηση σταθμών βάσης σε περιπτώσεις χαμηλής κίνησης στο δίκτυο. Βέβαια τα αποτελέσματα έδειξαν ότι σε κάποιες περιπτώσεις απενεργοποίησης σταθμών βάσης, οι τερματικοί σταθμοί ύστερα από τη μεταπομπή τους, συνδέθηκαν σε μακρινούς σταθμούς βάσης με αποτέλεσμα να αυξήσουν την καταναλισκόμενη ενέργεια. Σε αυτή την περίπτωση είναι εμφανής η χρησιμότητα του αλγορίθμου για τη δημιουργία συντομότερο μονοπατιών με την χρήση των τερματικών ως διαμεσολαβητών. Σε αρκετές περιπτώσεις περαιτέρω βελτίωση στην καταναλισκόμενη ενέργεια και ελάττωση των διακαναλικών παρεμβολών μπορεί να επιτευχθεί από την ανακατανομή των χρησιμοποιούμενων καναλιών. Επίσης, τα αποτελέσματα ανέδειξαν κάτι πολύ σημαντικό. Την πιθανότητα εμφάνισης συγκρούσεων μεταξύ των προβλημάτων ή των αναδιαμορφώσεων που ενεργοποιούνται ταυτόχρονα ή σχεδόν ταυτόχρονα. Το τελευταίο οδήγησε στην πρόταση ενός μηχανισμού για τον εντοπισμό και την αποφυγή συγκρούσεων ή αρνητικών αλληλεπιδράσεων μεταξύ των παραμέτρων ή των υπό παρακολούθηση μετρικών.

Ο προτεινόμενος μηχανισμός εντοπισμού και αποφυγής συγκρούσεων ή αλληλοεπιδράσεων μεταξύ των διαφόρων αλγορίθμων αναδιαμόρφωσης δέχεται ως είσοδο το σύνολο των μετρικών και των αναδιαμορφώσεων που ενεργοποιούνται. Στόχος του μηχανισμού είναι να επιλεγεί εκείνη η αναδιαμόρφωση η οποία θα επηρεάσει λιγότερο τη σταθερότητα του αυτο-οργανωμένου δικτύου, λαμβάνοντας υπόψη την κρισιμότητα και τη προτεραιότητα των μετρικών απόδοσης που ξεπερνούν τα προκαθορισμένα από το διαχειριστή του δικτύου κατώφλια. Ο σχετικός μηχανισμός αποτελείται από τρεις φάσεις. Στην πρώτη φάση χρησιμοποιείται η μέθοδος της διπλής εκθετικής εξομάλυνσης ώστε να αποφευχθούν αναδιαμορφώσεις που οφείλονται σε συνεχή διακύμανση ή παροδικότητα των συσχετιζόμενων μεταβλητών. Στη δεύτερη φάση εντοπίζονται συγκρούσεις μεταξύ των υποψήφιων αναδιαμορφώσεων και επιλύονται, ενώ στη τρίτη φάση αξιολογούνται οι μη συγκρουόμενες αναδιαμορφώσεις σχετικά με την επίδραση που έχουν στις υπόλοιπες μετρικές απόδοσης. Επίσης, προτείνεται ένα πιθανοτικό μοντέλο, όπου υπολογίζεται ο βαθμός εμπιστοσύνης για κάθε υποψήφια αναδιαμόρφωση και παίρνει προτεραιότητα να εκτελεστεί αυτή με τη μεγαλύτερη θετική επίδραση ή αλλιώς τη λιγότερο αρνητική στο σύνολο των μετρικών απόδοσης του δικτύου. Τα πειραματικά αποτελέσματα σε αυτή την περίπτωση δείχνουν ότι με το συγκεκριμένο μηχανισμό διασφαλίζεται η σταθερότητα σε ένα αυτο-διαχειριζόμενο δίκτυο, ενώ βελτιώνεται η συνολική απόδοση του δικτύου σύμφωνα με τις καθορισμένες προτεραιότητες των διαθέσιμων μετρικών.

Η συγκεκριμένη διατριβή περιλαμβάνει οχτώ κεφάλαια. Το κεφάλαιο 1 αποτελεί την εισαγωγή στο αντικείμενο της διατριβής, στην οποία αποτυπώνονται οι τεχνολογικές εξελίξεις στην περιοχή των δικτυακών υποδομών και υπηρεσιών στο πλαίσιο του μελλοντικού διαδικτύου. Περιγράφονται οι προκλήσεις στη διαχείριση των δικτυακών και ενεργειακών πόρων, εστιάζοντας σε ασύρματα δίκτυα τα οποία βρίσκονται σε αστικές περιοχές (γραφεία, οικίες κλπ) και σχηματίζουν πυκνές τοπολογίες με έντονη επικάλυψη τόσο γεωγραφική όσο και συχνοτήτων. Στη συνέχεια, αναφέρονται τα κίνητρα για την εισαγωγή των γνωσιακών και αυτο-οργανωμένων (αυτο-διαχειριζομένων) τηλεπικοινωνιακών συστημάτων. Καθορίζονται τα δικτυακά προβλήματα που μελετήθηκαν και συνοψίζεται η συνεισφορά της διατριβής σε επίπεδο μηχανισμών, αλγορίθμων και αποτελεσμάτων για την αποδοτικότερη χρήση των πόρων ασύρματων δικτύων και τη μείωση της ενέργειας. Στο τέλος του κεφαλαίου περιγράφεται η δομή της διατριβής.

Στο κεφάλαιο 2 γίνεται μια σύντομη εισαγωγή στις βασικές έννοιες της διατριβής. Συγκεκριμένα, ορίζονται και αναλύονται τα γνωσιακά και αυτο-οργανωμένα συστήματα επικοινωνιών, εστιάζοντας στα βασικά τους χαρακτηριστικά, στις ιδιότητές τους αλλά και τις κύριες απαιτήσεις για αυτά τα συστήματα. Ιδιαίτερη έμφαση δίνεται στις αρχιτεκτονικές απαιτήσεις και τις αντίστοιχες ικανότητες για την αποτελεσματική και αξιόπιστη λειτουργία των αυτο-διαχειριζόμενων συστημάτων. Επίσης, γίνεται μια επισκόπηση της σύγχρονης βιβλιογραφίας όσον αφορά τα αυτο-οργανώσιμα συστήματα επικοινωνιών και γίνεται μια σύντομη αναφορά σε αντιπροσωπευτικές προτάσεις αυτο-οργάνωσης σε διάφορους τομείς που εμπίπτουν στο αντικείμενο της διδακτορικής διατριβής. Στη συνέχεια, παρουσιάζονται διάφορες διαχειριστικές λειτουργίες που αφορούν τοπικά ασύρματα δίκτυα π.χ. IEEE 802.11, των οποίων η απόδοση μπορεί να βελτιωθεί από την εισαγωγή μηχανισμών αυτο-οργάνωσης και την ενίσχυση των γνωστικών τους δυνατοτήτων. Τέλος, γίνεται επισκόπηση των πρόσφατων μηχανισμών που υπάρχουν στη βιβλιογραφία για τη βέλτιστη χρήση των δικτυακών και ενεργειακών πόρων.

Στο κεφάλαιο 3, παρουσιάζεται το προτεινόμενο αρχιτεκτονικό πλαίσιο για την υλοποίηση των αυτο-οργανωμένων τηλεπικοινωνιακών συστημάτων. Συγκεκριμένα, εισάγεται ο γνωσιακός διαχειριστής δικτυακών στοιχείων (επίπεδο I) καθώς και ο γνωσιακός διαχειριστής δικτυακού τομέα (επίπεδο II), των οποίων οι λειτουργίες αναλύονται. Πρόκειται για πράκτορες λογισμικού οι οποίοι ακολουθούν τις βασικές αρχές των βρόχων ανάδρασης. Επίσης, δίδονται παραδείγματα από την υλοποίηση των παραπάνω πρακτόρων σε σταθμούς βάσης τοπικών ασύρματων δικτύων και αναφέρονται οι απαραίτητες διεπαφές με τις υπόλοιπες λειτουργίες και μηχανισμούς ενός σταθμού βάσης. Εν συνεχεία, περιγράφεται η γνωσιακή βάση των γνωσιακών διαχειριστών όπου αναπαριστώνται οι συσχετίσεις μεταξύ των μετρικών απόδοσης, των γεγονότων, των σφαλμάτων και των δράσεων αναδιαμόρφωσης ή προσαρμογής που υποστηρίζονται τους δικτυακούς κόμβους. Όλες αυτές οι συσχετίσεις χρησιμοποιούνται για το συμπερασμό και τη λήψη αποφάσεων που αφορούν τη διαχείριση των δικτυακών πόρων, ενώ έχουν επικουρικό ρόλο στο μηχανισμό αποφυγής συγκρούσεων που περιγράφεται σε παρακάτω κεφάλαιο. Τέλος, στο συγκεκριμένο κεφάλαιο σχολιάζονται πειραματικά αποτελέσματα σχετικά με την επιβάρυνση που προκαλούν οι γνωσιακοί διαχειριστές στις δικτυακές συσκευές όπου τοποθετούνται (π.χ. σταθμοί βάσης) σε επίπεδο CPU, RAM, και δικτυακού φόρτου.

Στο κεφάλαιο 4 περιγράφεται ο αλγόριθμος για τη δημιουργία συστάδων μεταξύ των κόμβων ενός αυτο-οργανωμένου ασύρματου τοπικού δικτύου και την εκλογή των αρχηγών (γνωσιακός διαχειριστής επιπέδου II) ανά συστάδα. Ο συγκεκριμένος αλγόριθμος επιτυγχάνει το σχηματισμό των συνεργατικών δομών (ομάδων) μεταξύ των σταθμών βάσης, με σκοπό την ανάπτυξη ευρύτερων αυτο-οργανωμένων σχηματισμών για την παρακολούθηση του δικτύου και τη λήψη αποφάσεων σχετικά με τη διαχείριση των πόρων. Ενώ σημαντικός είναι ο ρόλος τους στην εξομάλυνση συγκρούσεων από τους διάφορους μηχανισμούς αυτο-βελτίωσης των γειτονικών κόμβων. Αρχικά, γίνεται μια σύντομη επισκόπηση των διαθέσιμων αλγορίθμων συσταδοποίησης, από τις περιοχές των δικτύων ασύρματων αισθητήρων και των επί τούτω δικτύων, όπου αναφέρονται χαρακτηριστικά που μπορούν να χρησιμοποιηθούν αλλά και πιθανές δυσκολίες που εμποδίζουν την υιοθέτησή τους στην περιοχή των αυτο-οργανωμένων δικτύων. Ο προτεινόμενος κατανεμημένος αλγόριθμος βασιζόμενος στα τοπολογικά χαρακτηριστικά ή αλλιώς στα στοιχεία του σχηματιζόμενου γράφου της περιοχής του δικτύου (βαθμός των σταθμών βάσης, διάμετρος της περιοχής του δικτύου), και τον κανόνα της “κατά προτίμησης προσκόλλησης”, εκλέγει τους επικεφαλής των ομάδων και καθορίζει τα σύνορα της κάθε συστάδας. Περιγράφεται τόσο η φάση ανάπτυξης όσο και η φάση συντήρησης των σχηματιζόμενων συστάδων, εξαιτίας των αλλαγών στη δικτυακή τοπολογία. Επίσης, παρουσιάζονται και αναλύονται πειραματικά αποτελέσματα προσομοιώσεων για τη συμπεριφορά του προτεινόμενου αλγορίθμου, χρησιμοποιώντας διάφορες τοπολογίες δικτύου όσον αφορά τον αριθμό και την πυκνότητα των σταθμών βάσης. Τα αποτελέσματα δείχνουν την αποτελεσματική ανακάλυψη των συστάδων καθώς και το καλό επίπεδο modularity (Q) που επιτυγχάνεται στις σχηματισμένες συστάδες.

Στο κεφάλαιο 5 περιγράφονται οι αλγόριθμοι που έχουν προταθεί για τη βελτίωση της κάλυψης, της χωρητικότητας και των παρεμβολών ενός ασύρματου δικτύου καθώς και για την ελάττωση στην κατανάλωση της ενέργειας, αξιοποιώντας τους προαναφερθέντες μηχανισμούς των αυτο-οργανωμένων δικτύων. Αναλύεται η μέθοδος που χρησιμοποιήθηκε στη διατριβή για τον υπολογισμό της κατανάλωσης ενέργειας κατά τη μετάδοση πακέτων, καθώς και οι αντίστοιχες μετρικές για τον υπολογισμό της κάλυψη και της χωρητικότητας ενός ασύρματου δικτύου. Χρησιμοποιώντας ένα σχήμα βασισμένο σε κανόνες επιλέγεται ο κατάλληλος μηχανισμός αναδιαμόρφωσης του τοπικού δικτύου, αξιολογώντας τις μετρικές απόδοσης, τους περιορισμούς σε πόρους ή ενέργεια και τους κανόνες ή κατώφλια που έχει θέσει ο διαχειριστής του δικτύου. Ακολουθεί η περιγραφή των προτεινόμενων μηχανισμών για την αναδιαμόρφωση του τοπικού ασύρματου δικτύου. Αρχικά, προτείνεται ένας καινοτόμος αλγόριθμος για τη δυναμική ενεργοποίηση και απενεργοποίηση σταθμών βάσης που αποσκοπεί στην εξορθολογισμένη χρήση των ραδιο-πόρων βάσει της πυκνότητας και της κίνησης του δικτύου. Επίσης, παρουσιάζεται το σχήμα για τη μεταπομπή των τερματικών, ύστερα από κάθε ενεργοποίηση ή απενεργοποίηση σταθμού βάσης για την καλύτερη εξισορρόπηση του φόρτου. Στη συνέχεια περιγράφεται ο προτεινόμενος αλγόριθμος για το σχηματισμό μονοπατιών που αποτελούνται από πολλά άλματα, έχοντας ως ενδιάμεσους κόμβους (ή αναμεταδότες) τερματικές συσκευές (συσκευές χρηστών). Ο σκοπός του σχήματος είναι η ελάττωση της κατανάλωσης ενέργειας στην πλευρά των τερματικών συσκευών, αξιοποιώντας τοπικές ευκαιρίες δικτύωσης, και αντικαθιστώντας την απευθείας σύνδεση ενός τερματικού με ένα σταθμό βάσης, με μια συντομότερη διαδρομή που έχει έναν ή περισσότερους ενδιάμεσους αναμεταδότες (δλδ τερματικούς σταθμούς). Επίσης, εισάγεται ένα σχήμα επιλογής καναλιών/συχνοτήτων σύμφωνα με το φόρτο κάθε σταθμού βάσης. Στο τέλος του κεφαλαίου, παρουσιάζονται τα αποτελέσματα της προσομοίωσης που διενεργήθηκαν στον προσομοιωτή OPNET για τη συμπεριφορά των προτεινόμενων αλγορίθμων και τη βελτίωση στην κατανάλωση ενέργειας και στην αποδοτικότερη χρήση των ραδιο-πόρων σε ένα ασύρματο τοπικό δίκτυο τεχνολογίας IEEE 802.11. Συγκεκριμένα, για την τοπολογία που επιλέχθηκε η απενεργοποίηση σταθμών βάσης βελτιώνει την κατανάλωση για το τμήμα της λήψης 12%. Αντίθετα η κατανάλωση στην πλευρά των συσκευών του χρήστη αυξάνεται. Σε αυτή την περίπτωση η ανακατανομή καναλιών επιτυγχάνει την ελάττωση στην κατανάλωση για την λήψη δεδομένων τόσο στην πλευρά του δικτύου όσο και σε αυτή των χρηστών, 40.5% και 35.7%. Ο σχηματισμός μονοπατιών μεταξύ κάποιων συσκευών του χρήστη οδηγεί στην ελάττωση 20% της μέσης κατανάλωσης ενέργειας για τη μετάδοση δεδομένων από το χρήστη.

Στο κεφάλαιο 6 αναλύονται οι αλληλεπιδράσεις μεταξύ των διαφορετικών μηχανισμών αναδιαμόρφωσης ενός αυτο-οργανωμένου δικτύου και προτείνεται ο μηχανισμός αποφυγής συγκρούσεων. Όπως δείχνουν και τα πειραματικά αποτελέσματα ένα πρόβλημα που προκύπτει σε πολλές περιπτώσεις αυτο-βελτίωσης είναι τα συγκρουόμενα «συμφέροντα» ή στόχοι ακόμη και για τον ίδιο κόμβο, όπως επίσης και το ενδεχόμενο η βελτιστοποίηση μιας δικτυακής παραμέτρου να οδηγήσει στη συνέχεια σε ενεργοποίηση μιας σειράς μη επιθυμητών αναδιαμορφώσεων. Το γεγονός αυτό οδήγησε στην περιγραφή ενός μηχανισμού εντοπισμού και επίλυσης συγκρούσεων, καθώς και εξαρτήσεων, μεταξύ διαφορετικών μετρικών απόδοσης ή μεταξύ πολλαπλών ενεργειών αυτο-βελτίωσης που προέκυψαν από τη συλλογιστική μηχανή ενός αυτο-διαχειριζόμενου κόμβου. Ο προτεινόμενος μηχανισμός αποτελείται από τρεις φάσεις και καλείται κατά τη φάση της λήψης αποφάσεων ενός αυτο-οργανωμένου κόμβου (ή σχηματισμού) και αφού έχουν προσδιορισθεί οι απαραίτητες αναδιαμορφώσεις. Στην πρώτη φάση χρησιμοποιείται η μέθοδος της διπλής εκθετικής εξομάλυνσης από τις χρονολογικές σειρές ώστε να εντοπισθούν μετρικές οι οποίες παρουσιάζουν συνεχή διακύμανση ή παροδικότητα και να αποκλειστούν οι αναδιαμορφώσεις οι οποίες ενεργοποιούνται από τις αντίστοιχες μετρικές. Στην επόμενη φάση ελέγχονται οι επιλεγμένες από την πρώτη φάση αναδιαμορφώσεις για πιθανές συγκρούσεις. Δηλαδή κατά πόσο δύο ή περισσότερες μετρικές ζητούν αντίθετη (ως προς την κατεύθυνση) αναδιαμόρφωση πάνω στην ίδια παράμετρο. Η προτεραιότητα και η ακριβής κατάσταση της μετρικής που ενεργοποιεί μια αναδιαμόρφωση είναι τα κριτήρια επίλυσης της σύγκρουσης, καθορίζοντας την κατεύθυνση που τελικά θα επιλεγεί. Το αποτέλεσμα της δεύτερης φάσης και συγκεκριμένα εκείνες οι αναδιαμορφώσεις οι οποίες δεν παρουσιάζουν κάποια σύγκρουση, δίδεται ως είσοδος στη τρίτη φάση. Εδώ ελέγχονται οι συγκρούσεις στην πλευρά των μετρικών καθώς και η αλληλεπίδραση στις υπόλοιπες μετρικές απόδοσης. Επιπλέον, εισάγεται ένα πιθανοτικό σχήμα όπου υπολογίζεται ο συντελεστής βεβαιότητας για την εκτίμηση του αποτελέσματος κάθε διαθέσιμης αναδιαμόρφωσης. Συνολικά ο στόχος είναι η επιλογή της εκείνης της αναδιαμόρφωσης η οποία έχει τη μεγαλύτερη προτεραιότητα, δημιουργεί τις λιγότερες συγκρούσεις και έχει τις λιγότερες πιθανότητες να οδηγήσει σε χειροτέρευση των υπολοίπων μετρικών απόδοσης. Στη συνέχεια του κεφαλαίου η προτεινόμενη λύση για το συντονισμό των δράσεων ενός αυτο-οργανωμένου δικτύου εφαρμόζεται σε ένα ασύρματο τοπικό δίκτυο και στους αλγορίθμους που προτάθηκαν για τη βελτιστοποίηση της κάλυψης και της χωρητικότητας ενός ασύρματου δικτύου και για την ελάττωση στην κατανάλωση της ενέργειας. Τέλος, στο κεφάλαιο 6, παρουσιάζονται τα πειράματα που διενεργήθηκαν με τη χρήση του προσομοιωτή OPNET και αναλύονται τα αποτελέσματα για τη συμβολή του αλγόριθμο αποφυγής συγκρούσεων και επίλυσης αλληλεπιδράσεων στη σταθερότητα του αυτο-οργανωμένου δικτύου, στην αποδοτικότερη διαχείριση των ραδιο-πόρων και στην κατανάλωση ενέργειας.

Στο κεφάλαιο 7 περιγράφεται ένα αλγοριθμικό πλαίσιο για την επέκταση των γνωσιακών ικανοτήτων των διαχειριστών της προτεινόμενης αρχιτεκτονικής, με στόχο τη βελτίωση των διαδικασιών για τη διαχείριση της απόδοσης του δικτύου. Αυτό το πλαίσιο αποτελείται από τις φάσεις για τη λήψη αποφάσεων, τη μάθηση και την εκτέλεση των αναδιαμορφώσεων. Στη συγκεκριμένη περίπτωση εφαρμόζεται και μελετάται για τη βελτίωση της ποιότητας της υπηρεσίας VoIP σε ένα ασύρματο δίκτυο WiMAX, το οποίο παρουσιάζει συνθήκες υψηλού φόρτου. Οι προτεινόμενοι αλγόριθμοι μπορούν να εφαρμοστούν σε οποιαδήποτε δικτυακό περιβάλλον πχ WLAN, καθώς και σε διάφορα δικτυακά προβλήματα. Συγκεκριμένα, εισάγεται ένα σχήμα λήψης απόφασης, για την επιλογή της πιο κατάλληλης αναδιαμόρφωσης για την αντιμετώπιση των συνθηκών συμφόρησης του δικτύου, και συγκεκριμένα την ελάττωση στις απώλειες πακέτων VoIP (<1%). Το σχήμα επιλέγει μεταξύ της αλλαγής προτεραιότητας των ροών στο σταθμό βάσης και της αλλαγής των αλγορίθμων συμπίεσης που χρησιμοποιούν οι ροές της υπηρεσίας φωνής (VoIP). Η ένταση κάθε αναδιαμόρφωσης υπολογίζεται με βάση το ιστορικό προηγούμενων αντίστοιχων ενεργειών, ενώ προτείνεται και μια ευρετική μέθοδος για περιπτώσεις όπου δεν υπάρχει το σχετικό ιστορικό ενεργειών. Επίσης, προτείνεται ο μηχανισμός μάθησης, βασιζόμενος στον αλγόριθμο K-means, για την αυτόματη ρύθμιση του μηχανισμού λήψης αποφάσεων βασιζόμενος σε προηγούμενες αποφάσεις. Τέλος, παρουσιάζονται και αναλύονται τα πειραματικά αποτελέσματα που συλλέχθηκαν κατά τη δοκιμή του προτεινόμενου αλγοριθμικού πλαισίου, το οποίο αναπτύχθηκε στην πανευρωπαϊκή πλατφόρμα πειραματισμού (FIRE Panlab) χρησιμοποιώντας ένα δίκτυο WiMAX network. Η λήψη αποφάσεων βασισμένη στο ιστορικό ελαττώνει τις επαναλήψεις και την μεταβατική περίοδο. Ενώ η μάθηση με ανατροφοδότηση βελτιώνει την ακρίβεια στη λήψη αποφάσεων, αποφεύγοντας αναποτελεσματικές αναδιαμορφώσεις.

Τέλος, στο κεφάλαιο 8 συνοψίζεται η ερευνητική συνεισφορά της διατριβής και αποτυπώνονται τα ζητήματα που αντιμετωπίστηκαν και οι δοθείσες λύσεις. Ως τελευταία πράξη της διατριβής, παρουσιάζονται οι σημαντικότερες προεκτάσεις των ερευνητικών θεμάτων που μελετήθηκαν κατά την εκπόνηση της διατριβής και αναγνωρίζονται οι ενδιαφέρουσες μελλοντικές επεκτάσεις.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

FOREWORD

This is a doctoral thesis submitted to the National and Kapodistrian University of Athens for the doctoral degree. This work has been carried out at the SCAN Group (http://scan.di.uoa.gr), Department of Informatics and Telecommunications at the National and Kapodistrian University of Athens, under the supervision of Prof. Lazaros Merakos and Lect. Nancy Alonistioti.

CHAPTER 1

INTRODUCTION

1.1 Future Networks Landscape

Starting from late 1960s and early 1970s the Internet achieved the interconnection and the cooperation of remote networks following a layered, distributed and packet-based architecture. These principles facilitated its fast extension providing scalability and robustness features. Gradually the interaction of remote communication networks lead to the development of novel end-user services (e.g., WWW, multimedia), which the last decade define the criteria for the evolution of communication networks technologies and specifically for wireless access networks.

In the coming years the communication networks have to deal with various types of advanced and composite services that will follow emerging societal needs with respect to e-health, smart cities, multimedia streams, machine-to-machine etc. The forecasts by telecom players predict a massive traffic growth, with predictions in the order of 1000 times larger global wireless-traffic volumes in 2020 than seen in 2010. This is mainly driven by increased usage of mobile multi-media services. Apart from the increase in the traffic there is also an explosion of the number of devices (e.g., smartphones, tablets, and sensors). It is estimated that by the end of 2012 the number of mobile-connected devices will exceed the number of people on earth and by 2016 there will be 1.4 mobile devices per capita [1].

On the other side, wireless network (e.g., WiMAX, LTE, WLAN) are constantly progressing and new technologies are integrating in order to address the increasing capacity and quality of service (QoS) needs. In the last decade, there has been a continuous increase in the number of wireless access point (AP) installations (e.g., IEEE 802.11) in private and public places to cope with user mobility and capacity requirements for emerging and future Internet services. To this end, the density of base stations (BS) or access points will continue to increase in areas with a large density of active users. Large numbers of femto-cells will be deployed to improve home and small office coverage and offload traffic from macro cells. A large installed base of WiFi AP (>500 million) will carry traffic, mainly indoors [2]. This future Internet environment could be characterized as the “seamless fabric of classic networks and networked objects” [3], which will be service and media rich, but also resources hungry and cost sensitive.

1.2 Motivation and Challenges

All access devices (e.g., WLAN APs) that are placed in a dense urban environment are often not part of the same administrative entity, and the configuration of their locations and operational features are not necessarily planned for the “network welfare”. This unstructured network environment results in dense AP topologies, especially in urban areas, with high coverage and frequency overlapping. The above in conjunction with users’ varying traffic volume and service requirements create on the one hand management overhead, while on the other hand, optimization opportunities arise for the avoidance of wasting network and energy resources. Contrary to common belief, information and communication technologies contribute a significant portion both to world energy consumption (2-4%) and environmental pollution (2-2.5% of greenhouse gas) [4]. Wireless network energy efficiency plays a primary role in reducing the impact of communication systems on energy consumption and environmental pollution [5]. Apart from an environmental responsibility, energy saving is important for the reduction of communication networks operational costs.

The heterogeneity of traffic flows and the spatio-temporal dynamics of the wireless networks complicate network planning and configuration. Existing network management tools have limited autonomic capabilities and their operation is mainly centralized with limited cooperation among involved entities; a fact that raises important scalability and complexity issues. Hence, we have to devise new ways to manage network and radio resources, in order to optimize different performance metrics (e.g., throughput, capacity, energy), considering the varying service conditions. Network operators are called to evolve Network Management Systems (NMSs) towards this direction, reducing also the high operational cost for network control and management. Network operators need more and more well qualified human resources in order to solve detected faults or to configure optimally network resources. Recent studies show that the contribution of Operational Expenditures (OPEX) on the management cost is constantly increasing comparing to the Capital Expenditures (CAPEX) [6].

Taking into consideration the explosion of the number of network devices and the exponential increase of the traffic volume we notice that the design criteria for beyond 4G systems [7] include requirements that were not present in previous technology roadmaps, which were mainly focusing on data rate, latency or coverage issues. Some of them are the following:

- Efficient utilisation of network resources,
- Energy saving,
- Management cost,
- Complexity.

The need to cope with complexity that derives from the interaction of hundreds or even thousands of network devices for the identification and realization of optimization opportunities calls for a more distributed and localized solution. A Self-Organizing Network (SON) is considered as one of the most promising approaches for the management of networks that operate in highly dynamic and dense environments [8]. The engineering of the cognitive cycle, known also as feedback control loop is a key functionality that is necessary for SONs and consequently in network management realization. The management agent for the instantiation of the cognitive cycle is placed either in network devices or in elements dedicated to network management tasks. Cognitive cycle monitors periodically their operational environment utilizing active or passive monitoring tools, in order to build its awareness as regards faults or performance issues of the network node. Taking into account the outcomes of the situation awareness phase, the available knowledge models as well as the configuration actions that a network node supports, the cognitive cycle selects and applies the most appropriate re-configuration. Thereinafter, learning and reasoning schemes could be used for the evolution of a cognitive cycle deduction capability.

SONs can offer advanced energy-saving capabilities, contributing to a greener network environment [9], [10]. Energy management, which lies under the self-optimization umbrella, is already an identified SON functionality in structured cellular environments, e.g., 3rd Generation Partnership Program (3GPP) [11]. Energy-saving in a self-organized Wireless Local Area Network (WLAN) environment is indirectly achieved through the goal of the best use of radio and network resources. To this end, energy management should be studied in conjunction with other configuration or optimization management functionalities. The energy behavior of different components, devices, or even functions is not independent and their operation affects each other's energy consumption. Hence, the quantification of the energy consumption of different types of devices, as well as of the components of the same device (e.g., processing, communication) is necessary to analyze the effect of the different configuration actions on WLAN energy consumption. Due to the importance of the problem, there is need to enhance radio resource optimization approaches by introducing energy-specific metrics and management tasks with the explicit goal of reducing energy consumption. As presented below, several effective mechanisms have been proposed for network resources efficient utilization and more recently for energy efficiency in wireless communication systems. However, most of them focus on specific network elements and little progress has been made in analyzing the interactions and the dependencies of various solutions in different type of nodes and components.

Addressing the various cognitive cycles in a SON as units that operate individually might lead to conflicts or ripple effects; thus they affect the stability of the network or even leading to the degradation of network performance. The adoption of a totally centralized solution might has the advantage that a cognitive cycle will have the global view of the network status and optimal decision could be taken, but on the other hand we have a single point of failure, where computational, scalability, and local optimization or local search issues arise. Cognitive cycles that are placed in network devices (e.g., access points, routers) could be modeled as a network of interacting units. In the context of SONs the performance metrics of the various configuration or optimization problems and the associated configuration actions have various interactions and dependencies [12]. The avoidance of conflicts either on the parameters (i.e., adaptations, re-configurations) or on the metrics and the joint assessment of the various SON problems is a challenging issue for SONs. A conflict in a control parameter (e.g., AP switch On/Off, antenna tilt) appears if two or more metrics, which exceed a predefined threshold, attempt to change the same control parameter towards different (i.e., conflicting) directions. On other hand, a conflict on a performance metric (e.g., interference, load) indicates that a triggered configuration action has a negative influence on another performance metric. This might lead to a sequence of adaptations creating a ripple effect situation. Hence, in a SON which controls various performance management tasks (Coverage and Capacity Optimization (CCO) and energy saving (ES)) it is necessary to investigate their interrelation in order to resolve the various types of conflicts in the context of the same or neighboring device and select the most appropriate configuration action, ensuring the stability of the network.

Another key requirement for the efficient performance of SONs is the fine tuning of problem solving schemes that a cognitive manager uses for performance optimizations. Most of them are based on predetermined schemes. The cognitive manager should have the capability for tuning of network parameters (e.g., thresholds for Packet Error Rate (PER), objective functions weight) by exploiting the feedback from previous events or from historic data. The administrator of a network system (self-organizing or not) usually sets the initial values of these parameters or thresholds based on accumulated experience. However, the situation and the different characteristics of the network segments may lead to different values. Even if the initial selection of parameters per network node is correct, continuous updates during their digital life might be required. Feedback-based evolution of a SON could further enhance its performance management capabilities.

Taking into consideration the above analysis, we enlist below the research questions that this dissertation attempts to address:

- How in-network management is realized in a wireless network environment?
- How network nodes are coordinated in a self-organized network?
- How energy saving and efficient network resources utilization are achieved in a dense urban environment?
- How to resolve conflicts and dependencies among SON problems e.g., CCO, ES?
- How cognitive capabilities could improve the performance of a SON via learning techniques and history-based problem solvers?

1.3 General Framework and Dissertation Contribution

In this dissertation, we propose a novel approach for energy saving and wireless resources management in a WLAN urban environment, where dependencies among different types of nodes and components are taken into account. CCO adapts network connectivity at all desired locations and provides bandwidth according to the communication needs of the users, avoiding the overutilization and underutilization of network resources. Energy consumption is measured at both the AP and the User Equipment (UE) side, focusing on the communication component (transmission, reception). CCO is applied via a novel scheme for the dynamic deactivation or reactivation of APs. This scheme aims at the rational usage of the radio resources according to traffic intensity and network density. The mechanism for UE load balancing after the de(re)-activation of an AP is also provided. The effect of an AP deactivation on UE and other APs energy consumption is assessed, triggering an additional adaptation action in the case that an energy efficiency problem has been detected. A scheme for multi-hop relay communication mode is proposed for the energy saving of UE transmission phase, exploiting local networking opportunities. In addition, a novel channel re-allocation scheme is introduced for the reduction of energy consumption during data reception phase. CCO and energy saving algorithms in a SON WLAN have been evaluated using a simulation environment based on the OPNET simulation tool.

For the deployment of a SON, each AP incorporates a Cognitive Network Manager (CNM), where energy saving and CCO algorithms are placed. In the context of this dissertation we present two types of a CNM a) simple CNM that is referred to as Network Element Cognitive Manager (NECM) and b) domain CNM entitled Network Domain Cognitive Manager (NDCM). NECM implements the cognitive cycle at the network element level, providing an intelligent adaptation layer to the conventional control plane. Management problems that cannot be addressed directly at the network element level, due to computational or communicational constraints, are escalated to the respective NDCM level. The NDCM incorporates the required cognitive capabilities to identify optimization opportunities and solve problems that require a greater view of network status. The distributed software architecture of cognitive managers is described. We implemented the distributed cognitive framework (software agents and artificial intelligence algorithms) that is deployed in access points and base station of a real heterogeneous access network composed of a Broadband Worldwide Interoperability for Microwave Access (WiMAX) BS and WiFi AP. Interference management and load balancing through channel reselection and vertical assisted handover algorithms respectively are the management tasks of the NECM and NDCM in this experimentation phase. Useful findings and the recommendations from the deployment of the cognitive network management architecture in a real life implementation are provided (average utilization of processing resources, memory usage, and delay of cognitive cycle phases).

Central to our approach is the organization of WLAN APs into clusters to facilitate local management and coordination. Clusters are organization structures used for the collaborative tackling of network management problems, and they are formed following a common known scheme. Clusters facilitate the cooperation and the coordination of a group of network nodes for identifying and solving network management problems. In the literature, there are several clustering algorithms, which are mainly targeting wireless sensor networks or mobile ad-hoc networks, but the majority of them are application-specific (e.g., energy-efficient, mobility-aware). We propose SYSTAS algorithm, for the distributed discovery and establishment of clusters among network nodes, based on the features of the physical network topology. The density of the network graph and the preferential attachment model are used in order to form logical topologies i.e., clusters. The application of the proposed algorithm leads to the election of the head and the specification of the borders of the clusters through the allocation of the member nodes to the elected heads. The number of elected heads defines the number of the formed clusters. Clusters are non-overlapping and consist of two types of nodes:

- Simple member node,
- Head node.

The head node of each cluster has the role of a NDCM, while simple member nodes instantiate the NECM. Both types of nodes implement CCO and energy saving management tasks. The simulation results, using various network graphs, show the effective cluster formation and the resulted high modularity.

We propose a novel scheme for the management of the interactions of various optimization or configuration problems in a SON. We identify and resolve conflicts on metrics and parameters (i.e., configuration actions), which arise from the deduction phase of the cognitive managers that are placed in a SON, in the context of the same or neighboring devices. The proposed scheme consists of three steps. Firstly, a time series-based mechanisms is used in order to avoid checking a configuration action (adaptations) that is triggered by a performance metric, which value appears continuous variations due to temporary changes of the network area. This phase helps a SON to act proactively on conflicts and dependencies resolving, avoiding the trigger of adaptations that appear high uncertainty. In the next step, we check for conflicts on the triggered configurations actions. Only one “direction” per configuration action is prioritized, according to the severity and the priority of the performance metrics that trigger the corresponding configuration action. Finally, the impact of non-conflicting configuration actions on the other performance metrics is analyzed using a cost-benefit analysis scheme. The goal is to select the highest priority configuration action that creates fewer conflicts among available configuration actions and has the minimum possibility to deteriorate other (high priority) performance metrics. The proposed scheme for the coordination of various SON problems has been implemented and tested using OPNET simulation environment addressing CCO, energy saving, and interference tasks.

In addition, we describe an algorithmic framework for the extension of cognitive capabilities in network management, facilitating performance management tasks, and taking into consideration the system architecture that has been described in chapter 3. This framework is used for the improvement of Voice over IP (VoIP) QoS in a congested WiMAX network. Despite the WiMAX related introduction in the rest of this chapter the proposed algorithmic framework solution is not access technology specific, but is equally feasible to other wireless network technologies as well, such as WLAN. The proposed algorithmic framework consists of the decision making, the execution and the learning phase. The decision making part includes the scheme for the identification of the most appropriate action for the packet loss reduction of VoIP service; selecting between a) the change of VoIP flows priority at the WiMAX base station, exploiting Medium Access Control (MAC) features, and b) the change of VoIP flows selected codec, exploiting service level features. The solution and the quantification of the derived action (e.g., number of VoIP flows, the type of codec transition) is achieved using either a history-based scheme that takes advantage of previous events, or a heuristic approach for un-classified (i.e., unknown) situations. A k-Means learning algorithm is introduced to process the accumulated knowledge from all applied actions and evolve the decision making scheme. The performance and feasibility evaluation of the proposed solution has been tested using FIRE Panlab WiMAX experimental facility.

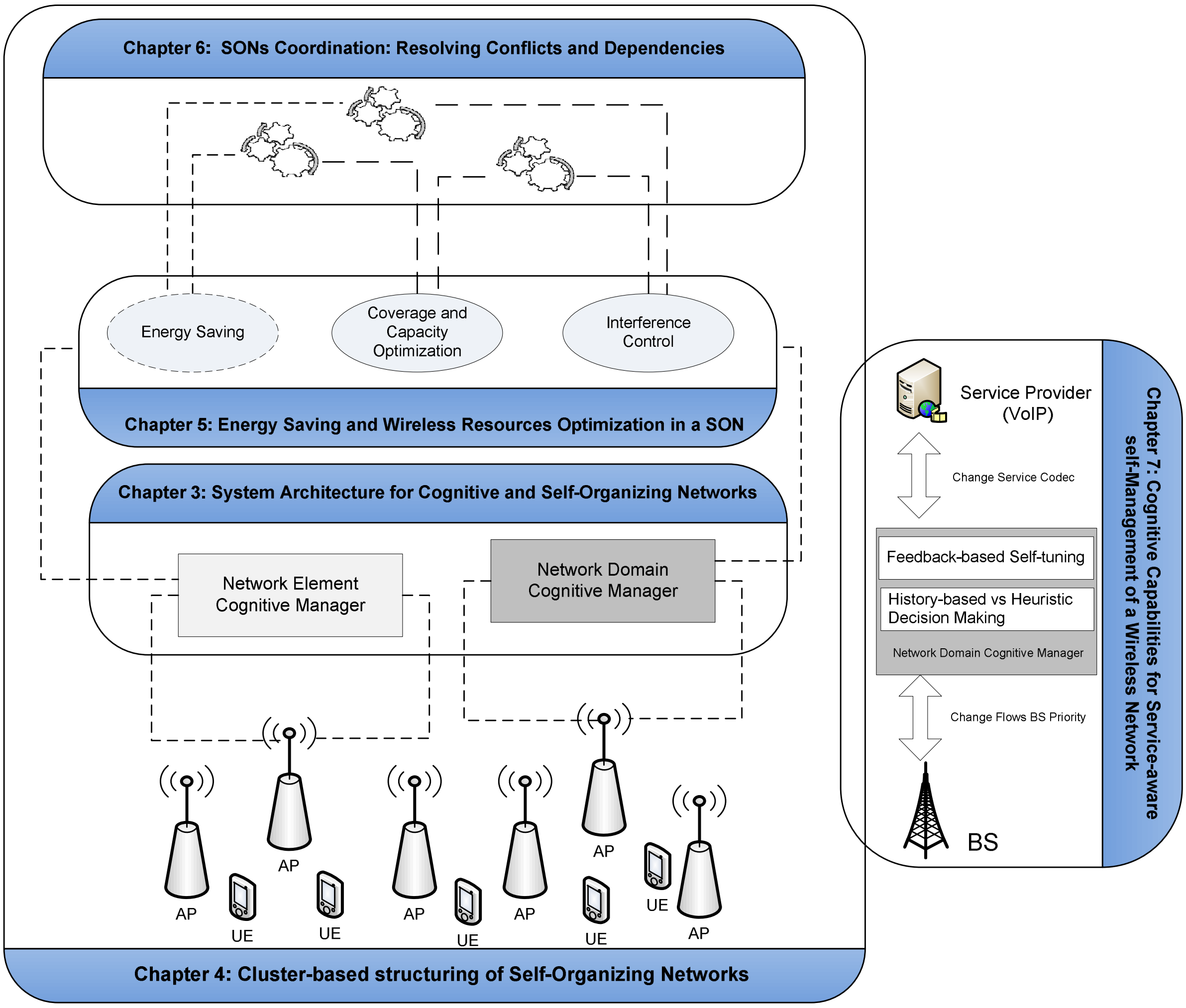


Figure 1-1: Schematic overview of the contributions of this dissertation - Different chapters focus is highlighted

The main contributions of the dissertation are summarized as follows:

- A novel approach for wireless resource management and energy saving at both user equipment and access point side in a self-organized WLAN urban environment. The proposed algorithms include: a) AP dynamic de/re-activation, b) UE multi-hop relays formation, and c) AP frequency channel re-allocation.
- Functional and software architecture for the development of a self-organized wireless network environment (Network Element Cognitive Manager and Network Domain Cognitive Manager).
- SYSTAS algorithm for the distributed discovery and formation of clusters of access points in a WLAN environment, based on the topological characteristics of the network graph. A cluster-based SON facilitates network nodes (cognitive managers) cooperation and coordination for CCO and energy saving management tasks.
- A scheme for the resolution of dependencies and conflicts among multiple SON problems (energy saving, CCO, interference).
- An algorithmic framework that extends the cognitive capabilities of a SON, using: a) history data of previous adaptations, and b) a feedback-based learning algorithm to process the accumulated knowledge in order to evolve the decision making scheme. This framework strengthens the accuracy of decision making, avoiding iterations or adaptations that are not effective.

1.4 Dissertation Structure

The dissertation is structured into eight chapters. Following this chapter, the structure is briefly presented below:

Chapter 2 introduces the main concepts of the dissertation. Specifically cognitive and self-organizing networks are defined and analyzed, focusing on the main features, properties and requirements of these systems. Various architectures for cognitive and autonomic communication networks are presented, while the related work for development of self-organizing communication networks is described. Management tasks for wireless networks (e.g., IEEE 802 11 WLANs), which could be enhanced introducing cognitive and self-organizing capabilities are enumerated. An overview of the literature work on coverage and capacity optimization and energy efficiency in wireless communications is provided as well as the background work for the interactions of SONs optimization problems.

In **Chapter 3** the functional and software architecture of the cognitive framework for self-organizing networks are presented. The deployment and assessment of the framework (network element cognitive manager-NECM, network domain cognitive manager-NDCM) on WiFi/WiMAX wireless environment are explained. Finally, results and the lessons learned from the experimentations activities are discussed and recommendations are proposed in this chapter.

Chapter 4 provides an overview of various clustering algorithms that have been proposed in mobile-ad-hoc and sensor networks, highlighting the categories and the properties a cluster. SYSTAS algorithm is proposed for the distributed discovery and establishment of clusters among network nodes (i.e., APs in a wireless access topology), based on the features of the physical network topology. The density of the network graph and the preferential attachment model are used in order to form logical topologies (i.e., clusters). The effectiveness of the proposed algorithm, in terms of the achieved modularity, is evaluated using various network topology graphs.

In **Chapter 5** the management functions for efficient utilization of network resources and energy saving are presented in a cluster-based self-organized wireless network. The metrics that are used for the calculation of energy consumption as well as for coverage and capacity in wireless networks are described. A rule-based decision making scheme is proposed for the selection of the appropriate adaptation. Moreover, the algorithm for AP dynamic deactivation or reactivation, the scheme for user equipments topology re-organization, and the algorithm for frequency channels re-allocation are presented. Thereinafter, this chapter presents the obtained simulation results on the energy saving attained by the proposed schemes, and discusses various aspects of energy consumption in access networks and user equipment.

In **Chapter 6** the interactions among different self-optimization and self-configuration problems of a SON are analyzed and a novel scheme is introduced for the resolution of conflicts and dependencies in a wireless communication network. The proposed scheme is implemented using OPNET simulation environment, where conflicts and dependencies among CCO, energy saving and interference in a WLAN environment are studied. The results show that the SON coordination scheme facilitates the network-level performance improvement assessing the severity of detected problems, without proceeding to needless adaptations that affect the stability of the network or extend its transitory period.

Chapter 7 describes an algorithmic framework for the incorporation of cognitive capabilities in the agents of the proposed architecture, facilitating performance management tasks. This framework, which includes the decision making the learning and the execution phases, is used for the improvement of VoIP QoS in a congested WiMAX network. Despite the WiMAX related introduction the proposed algorithmic framework solution is not access technology specific, but is equally feasible to other broadband wireless networks as well, such as WLAN. A decision making scheme for the selection of the most appropriate adaptation under congestion conditions, choosing between VoIP flows' priority change at the wireless base station and the change of VoIP flows' codec is introduced. A History-based method calculates the intensity of each adaptation, while a heuristic approach is used for un-classified situations. The proposed learning scheme, based on the feedback of previous actions, self-tunes the decision making tasks. Finally, we report the collected results from the proposed solution, which has been implemented and evaluated in FIRE Panlab experimental facility using a WiMAX network.

In **Chapter 8** the conclusions and the scientific contribution of the dissertation are summarized, by describing the challenges that have been addressed and the solutions that have been proposed. Open issues and suggestions for future work are also provided.

CHAPTER 2

TOWARDS A SELF-MANAGED FUTURE INTERNET

Literature investigates the introduction of cognitive, autonomous and self-organization capabilities in future communication systems so as to increase their intelligence, improve their performance, and handle the emerging complexity. Cognitive systems, autonomic communications as well as self-organization are interrelated scientific areas that are briefly surveyed in this chapter so as to identify functional requirements, recent advances, and proposed architectures, which attempt to increase the cognitive and adaptive behavior of a communication system. Thereinafter, wireless resources management in SON (energy saving, coverage and capacity optimization) and the interactions of different SON problems are studied.

2.1 Definitions

The definitions of the main concepts that are used in this chapter are summarized below:

- Network Management: concerns the operations used by the network to improve its performance and to define explicit policy rules for security, handling special customers, defining services, accounting, and so on. It also provides capabilities for monitoring the traffic and the state of network equipment. The philosophy of network management is that it should operate on a slow time scale and provide network elements with the information they need to react on faster time scales as the context dictates. Network management differs from signaling, since signaling mechanisms react to external causes on a very fast time scales and serve as the nervous response system of the network. Network management operations take place more slowly. They are triggered by the network administrator or control software detecting that some reallocation or expansion of resources is needed to serve the active contracts at the desired quality level. For example, when a link or a node fails, signaling is invoked first to choose a default alternative. At a later stage this decision is improved by the network management making an update to routing tables [13].
- Autonomic Network Management: management systems that are capable of self-governing and reducing the duties of the human operators who are not able to deal with increasingly complex situations. The systems should exhibit some level of intelligence so that their capability can improve over time, assuming more and more tasks that are initially allocated to skilled

administrators [5].

- Cognitive Network: a network with a cognitive process that can perceive current network conditions, plan, decide, act on those conditions, learn from the consequences of its actions, all while following end-to-end goals. This cognition loop, senses the environment, plans actions according to input from sensors and network policies, decides which scenario fits best its end-to-end purpose using a reasoning engine, and finally acts on the chosen. The system learns from the past (situations, plans, decisions, actions) and uses this knowledge to improve the decisions in the future [14].
- Self-Organizing System: We call a system organized if it has a certain structure and functionality. Structure means that the entities are arranged in a particular manner and interact (communicate) with each other in some way. Functionality means that the overall system fulfills a certain purpose. For example, a school of small fish tries to achieve a group structure that protects the fish against enemies. A system is self-organized if it is organized without any external or central dedicated control entity. In other words, the individual entities interact directly with each other in a distributed peer-to-peer fashion. Interaction between the entities is usually localized [15].
- Self-Organizing Network: can be defined as a set of use cases that govern a network including the planning, set up and maintenance activities. In this way the self-organizing networks enable the network to set itself up and then manage the resources to enable the optimum performance to be achieved at all times [16].

2.2 Background on Network Management

Several architectures have been proposed for network management the last decades. The Telecommunications Management Network (TMN) architecture is a reference model for a hierarchical telecommunications management approach. Its purpose is to partition the functional areas of management into layers. The International Telecommunication Union-ITU–T defined the TMN architecture in 1988 and it is described in Recommendation M.3010 [17] and other documents. Recommendation M.3010 defines the general TMN management concepts and introduces several management architectures at different levels of abstraction one of which defines a model that shows how management can be structured according to different responsibilities. This architecture is characterized by a hierarchical and layered organization of functions from the Business Management Layer to the network Element Management Layer. The figure below depicts this stacked organization. The Business Management layer is responsible for the management of the whole enterprise. This layer has a broad scope; communications management is just a part of it. Business management can be seen as goal setting, rather than goal achieving. The Service Management layer is concerned with management of those aspects that may directly be observed by the users of the telecommunication network. The responsibility of the Network Management layer is to manage the functions related to the interaction between multiple pieces of equipment. The functions of individual Network Elements are managed by Operations Systems Functions (OSF) in the Element Management layer. This layer deals with vendor specific management functions and hides these functions from the layer above, the Network Management layer. This architecture relies on northbound and southbound interfaces between each layer to ensure the information flow. Furthermore, in addition to the TMN–layering structure, the ITU–T also splits the general-management functionality offered by systems into the five key areas of fault, configuration, accounting, performance, and security (FCAPS). This categorization is a functional one.

In combination to this architecture, Internet Engineering Task Force (IETF) focused its efforts on the Simple Network Management Protocol (SNMP) [18] and Management Information Base (MIB) at the two key components of the overall Internet management framework. Implicit in the SNMP architectural model is a collection of network management stations and network elements. Network management stations execute management applications which monitor and control network elements. Network elements are devices such as hosts, gateways, terminal servers, and the like, which have management agents responsible for performing the network management functions requested by the network management stations. The SNMP is used to communicate management information between the network management stations and the agents in the network elements. Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB [19].

Thereinafter, more advanced approaches have started being integrated, such as software agents [20], active networks [21] and policy-based systems [22], [23] aiming at partially automating the management tasks in order to achieve faster response times and lower management costs. Nonetheless, NMSs kept struggling to cope with newly emerging issues due to the ever-increasing size and complexity of the underlying network components and services.

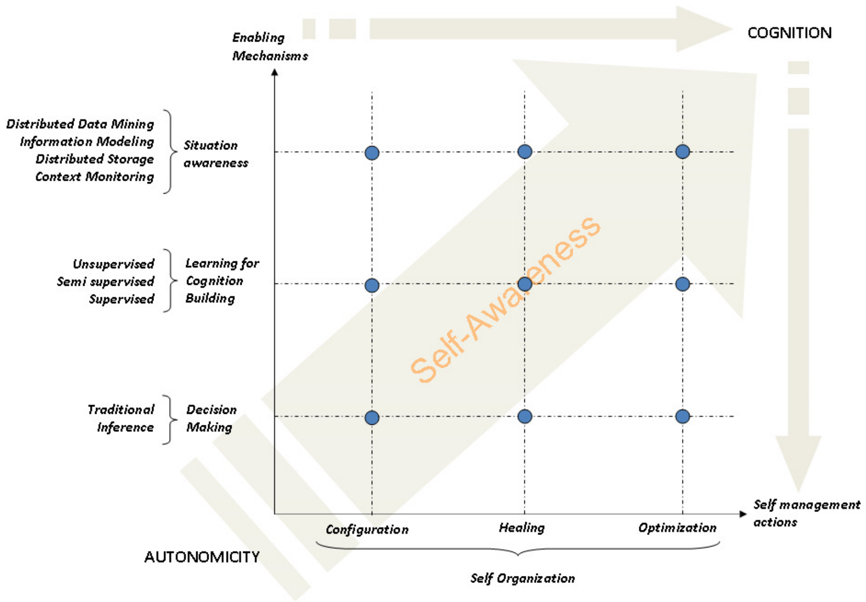


Figure 2-1: Conceptual model of future communication systems

Recently, the concepts of cognitive and self-organizing networks came to provide the paradigm shift that is required for network management. In the following sections we present recent literature work for cognitive and self-organizing networks. In many cases the terms autonomic network management and cognitive network management are used to describe similar functionality. Hence, we have selected to discuss both concepts in the same section. Figure 2-1 visualizes the gradual evolution from autonomy to cognition and self-organization. In this context, self-organization of network elements, at all scales, is considered to support collective and emergent cognition, which is amplified by multiple elements collaborative interaction. High level cognition is also attained by bottom-up organization.

2.3 Cognitive Network Systems

The fundamental features of an artificial cognitive system are embodiment, anticipation, adaptation, motivation, reasoning, and autonomy; these features are needed, since such architecture comprises a continue process of perception and action. A thorough review of artificial cognitive systems and cognitive architectures that is provided by Cliff [24], D. Vernon et al. [25], and P. Langley et al. [26] indicates various different architectures for cognitive systems, with various application fields such as robotics. These architectures are classified into cognitivist, emergent and hybrid models, taking into account their viewpoint on cognition and the different phases in cognitive science evolution. SOAR [27] and ACT-R [28] are two representative examples of the cognitivist approach, while Self-aware and Self-effective (SASE) architecture [29] follows the emergent systems approach. Several research initiatives have taken place recently trying to introduce cognition in communication network systems.

According to Thomas et al. [30], a cognitive network bears a cognitive process that can perceive current network conditions, and then plan, decide, and act on those conditions. The cognitive network can learn from these adaptations and use them to make future decisions. A framework is proposed in [30] to introduce cognition in the whole network taking into account end-to-end goals, and utilizing Software Adaptable Networks (SAN). Thomas et al. attempt to progress cognitive radio concept, which has been introduced by J. Mitola [31], by covering all aspects of communication networks, both wired and wireless/mobile. The introduction of cognitive capabilities in a communication system will continuously increase its intelligence, by viewing a problem in more than one ways, and by evolving its problem-solving process. In addition, the capitalization of cognition capabilities renders the system autonomous. This statement was firstly supported by Varela [32], who mentions that “cognition is an activity that facilitates the autonomy of a system by making coherent sense of environmental perturbations, so that the system’s autonomy is preserved”. In addition, Clark et al. [33] were the first to propose a new kind of network, which is aware of itself and its surroundings, thus a self-aware network, able to learn, decide and act according to those decisions to reach high-level goals, thus a network which employs cognition. Such network should, in the first phase, be able to recognize malfunctions and explain them in a (perhaps restricted) natural language. Then it should suggest ways of solving the malfunction and, finally, it should fix the problems itself.

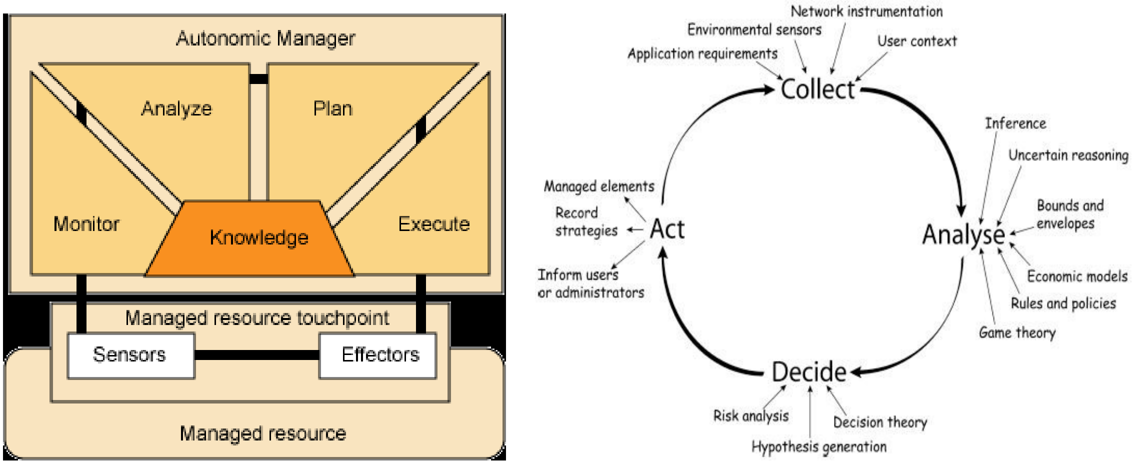
The challenge for the management systems of future Internet networks is the reduction of human intervention in the fundamental management functions and the development of the capabilities mechanisms that will render the network capable to configure, optimize, heal and protect itself, handling in parallel the emerging complexity. The IBM’s autonomic model [34] which was published in 2003, also known as self-CHOP model (self -configuration, -healing, -optimization and -protection), has been the milestone for the initiation of massive efforts towards self-managed systems. Although the above-mention paper is referred to the computing world, quickly the same principles were adopted to the communication networks for various control and management plane tasks. As it is stated by Quitadamo and Zambonelli [35] autonomic communications emphasizes more on distributed systems and services and to the management of network resources (e.g., spectrum) at both the user and infrastructure levels.

Autonomic network management indicates the necessity for communication networks engineering that are not static as regards their configuration, but can adapt to the changing demands placed upon them in open dynamic environments with reduced human interaction and steering, shifting away from systems which are developed according to a set of requirements agreed a priori [35], [36]. Their governance features is set by high-level policies that network nodes implement, by dynamically adapting network parameters (e.g., routing, protocols, queue sizes, radio frequencies) according to the varying network conditions (e.g., changes in topology, resource availability, traffic demands).

The majority of the existing proposals in the research literature for autonomic systems are based on the so called closed control loop (Figure 2-2). Each autonomic element consists of the autonomic manager (AM) and the respective managed resource. The AM monitors resources' state through the available sensors and builds the knowledge model that is used in conjunction with the monitoring data in order to analyze the current status of the managed resource and thereinafter decide or plan the best action. Cognition embodiment is an important aspect of future Internet self-managed systems, which complements and advances the automation of configuration actions. In-network cognitive cycles will allow communication systems to improve their inference and reasoning capabilities, by exploiting the feedback from previous events or from historic data that are stored locally. In the literature, there are several simple or more complex multidisciplinary models for cognition development e.g., [25], [37], [38]. However, the majority of them have not being designed considering the restrictions or capabilities that communication networks have. On the other hand, the cognitive models that refer to communication systems do not describe in many cases the available options and how they could be applied in a holistic networking context. Thus, there is the need to present how such ideas could be engineered in a real-world implementation.

Autonomic network elements are the elementary building blocks of autonomic systems, which produce the overall self-managing behavior through their mutual interactions leading to decentralized structures. Technologies associated with system-level self-configuration, self-healing, self-optimization, and self-protection entail interactions among multiple autonomic elements. The principle of decentralization is of utmost importance for autonomic systems, since any centralized resource or control point will impede a system's ability to adapt, especially in terms of robustness of performance.

Another challenge of self-managing systems is to achieve effective interoperation among autonomic elements [39]. Parashar [40] and Hariri et al. [41] define a conceptual architecture of an autonomic system based on the interrelated operation of the local and



(a) **(b)**

Figure 2-2: (a) IBM control loop (MAPE) [34], (b) Autonomic control loop [36]

global control loop. The former handles only the known environment states, while the global control loop manages the behavior of the overall system and defines the knowledge that drives the local adaptations.

Designing an autonomic network management system involves several technologies and disciplines and has received significant research effort the last decade. Several European research projects are working towards this direction. A non exhaustive list of architectures for autonomic management or communication networks self-managements are presented in Table 2-1.

Table 2-1: List of Architectures for Communication Networks Self-Management

F.I. Architecture	Description
4D [42]	It is a clean slate architecture that introduces four planes decision, dissemination, discovery, and data, focusing on core network. It focuses on IP networks.
4WARD [43]	It is a clean slate architecture that introduces In-Network Management (INM) functions that are located close to the management services, in most of the cases co-located on the same nodes. It focuses on various network environments (wireless, wired).
ANA [44]	It is a novel approach that can incorporate clean-slate or legacy solutions. ANA introduces an autonomic network meta-architecture that enables flexible, dynamic, and fully autonomous formation of network nodes as well as whole networks according to the working, economical and social needs of the users. It focuses on various network environments (wireless, wired).
AutoI [45]	It is a clean slate approach that introduces an architectural model consisting of a number of distributed management systems running within the network, which are described with the help of five abstractions and distributed systems: Virtualization, Management, Knowledge, Service Enablers and Orchestration Planes. It focuses on various network environments (wireless, wired).

SerWorks [46]	It is a clean slate approach that introduces an Consists of three frameworks: the Service Framework in the upper layer of the architecture, the Interaction Framework in the middle layer, and the Networking Framework in the lowest layer. It initially focused on Wireless Sensor Network (WSN) solutions but it is extended for more generic service infrastructure in wireless and also wired domain
CASCADAS [47]	It is a clean slate approach that introduces Autonomic Communication Element concept, which is an abstracted component model is used for situated and autonomic communication entities, at all levels of granularity. It focuses on various network environments (wireless, wired).
CONMan [48]	It is a clean slate approach that introduces protocol module abstractions (concepts, properties, capabilities) that enable manageability of future protocols in a “complexity-obvious way” and allow for dynamic protocol stack composition on the fly. It focuses on on various network environments (wireless, wired).
E3 [49]	It is an evolutionary approach that introduces Cognitive Management of heterogeneous networks wireless access part, exploiting local and global pilot channels. It focuses on heterogeneous wireless access networks environments.
FOCALE [50]	It is an incremental approach that can incorporate both clean-slate and legacy solutions. FOCALE introduces the architecture for network entities to self-govern their behavior within the constraints of business goals that the network as a whole seeks to achieve. It focuses on various network environments (wireless, wired).
GANA [51]	It is an incremental / evolutionary approach that introduces an architectural Reference Model of a Generic Autonomic Network Architecture that follows four levels of abstractions for which can be designed. It focuses on various network environments (wireless, wired).
Nestor [52]	It is an evolutionary approach that introduces an architecture for the automation of configuration by using policy scripts that access and manipulate respective network elements via a resource directory server (RDS). It combines several techniques from object modeling, constraint systems, active databases, and distributed systems. It focuses on IP network infrastructures.
Self-NET [53]	It is an evolutionary approach that introduces Cognitive network element and Cognitive domain managers for future Internet elements self-management in a semi-distributed and cooperative manner. If focuses on various network environments (wireless, wired).
Socrates [54]	It is an incremental approach that introduces Self-* functionalities for Self-Organizing networks. It focuses on 3GPP LTE network environments and specifically radio access side.

All these architectures that are presented in Table 2-1, which are also broadly reviewed in [55], [56], converge to a common agreement on the major functional blocks that are required for the specification of an autonomic network management system. Context-awareness, the existence of the knowledge plane, policy-based decision making and the network operator governance block constitute the main common components of the proposed architectures.

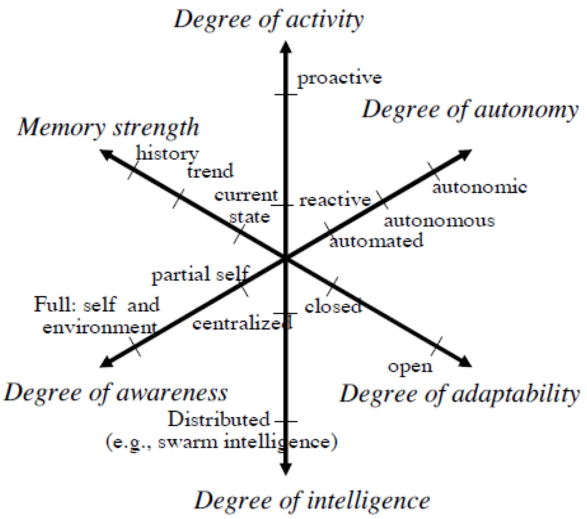


Figure 2-3: Classification dimensions of autonomic systems [56]

ETSI AFI has recently launched the specification of autonomic management systems [57], [58]. In addition some first implementation works have been initiated. AutoI project studied deployment and maintenance of end-users and networking services in fixed and wireless network environment with experiments based on virtual nodes [45]. INM prototyped in-network management functions in fixed networks and MANETs [59]. Autonomic Network Architecture (ANA) project developed a clean-slate meta-management architecture with inherent autonomic behaviors to flexibly host, interconnect and federate multiple heterogeneous networks [44]. The performed experiments, mainly in Mobile Ad-hoc Network (MANET) with virtual nodes, include scalability tests of the proposed unified address management framework, reliability and QoS performance tests and the efficiency of the proposed self-optimization mechanism. However, little progress has been made for the software engineering of the cognitive cycle, and the appropriate algorithmic schemes. In addition, little has been tried out on real networks experimentation and on the evaluation of feasibility or performance of cognitive management. The above mentioned issues are studied in Chapter 3, through software engineering and real deployment and evaluation of the cognitive management. Furthermore, recommendations and lessons learned are provided for the deployment of cognitive network management.

2.4 Self-Organizing Network Systems

Self-organization is another key feature of future communication systems, which can be viewed as a capability that complements adaptive behavior of communication systems and contributes towards their autonomy. Self-organized systems have the capability to change their organization without any external or central dedicated control entity. Self-organization goes beyond mere distribution and may not be based on global state information. Multiple individual entities interact in a distributed, collaborative peer-to-peer fashion (at a microscopic level) on a common global objective, which leads to sophisticated organization and defines the behavior of the global system (at a macroscopic level), thus establishing emergent[1] properties [15]. Self-organized systems are flexible, scalable, adaptive, robust to failures, and more reliable, since they degrade softly rather than break down suddenly.

The features of self-organized systems are necessary for future communications systems that operate in high dynamic and complex environments, considering the frequency of potential changes in their structure and their parameters. Mobile Ad Hoc Networks (MANETs) and Peer-to-Peer (P2P) networks are examples of dynamic network environments, where self-organization propertied and have been initially discussed [60]. Various algorithms and techniques have tried to solve networking issues for self-organized systems such as optimal path selection and service discovery [61]. Nature has been identified as one of the main sources of inspiration for self-organized systems providing many systems that could be studied.

Initially, self-organization techniques have been described for WSNs in order to minimize energy consumption, optimize data routing and data aggregation points [62]. Some of these approaches could be useful paradigms for next generation heterogeneous networks, taking always into account the specific limitations and characteristics that WSNs have such as low wireless communicational range, limited computational capabilities, predefined tasks that specialize the objective of the self-organization purpose.

One of the first attempts in the literature to introduce self-organization capabilities in wireless communications systems, to the best of our knowledge, was initiated by Baker and Ephremides [63]. The authors propose the linked cluster algorithm (LCA), which enables a specific network type (i.e., HF ITF) to remodel itself and maintain balanced connectivities in a self-organizing manner without the existence of a central controller. Robertazzi and Sarachik [64] discuss the early stages of the self-organizing communication systems design, focusing on the LCA algorithm and how topological and traffic information is distributed in an efficient manner, using the distributed evolutionary

---

[1] The term emergent properties (or emergence) in this paper describes the design paradigms, where behavior is not strictly programmed but robustly emerges from the interaction of the various components.

algorithm (DEA). An adaptive clustering algorithm for use in a self-organizing communication network, which recommends cluster heads and (re-) form connected clusters automatically, using the number of neighbors of a node as the main criterion is proposed in [65].

The last decade, applications of self-organized cellular communication systems have gained more attention by the research community. Self-organizing dynamic channel allocation, based on a local measurement of Carrier-to-Interference (C/I) at both the mobile nodes and base stations are discussed in [66] by Spilling and Nix. A joint cellular network and ad-hoc wireless network is considered in [67], using fixed relay stations in order to extend the coverage and QoS capabilities of cellular systems. The authors investigate the necessary conditions and parameter values that will allow nodes efficient distributed collaboration (self-organization) and they formulate the topology generation problem as a constrained optimization technique. Jiang et al. [68] describe a self-organizing approach for relay-based cellular station. Base stations and relay stations cooperate in order to ensure that the capacity is efficiently allocated as well as there is no gaps in the coverage, while a centralized optimization point is selected something that looses the self-organization principles [69]. Finally, Lu et al. present in [70] a framework for the organization of a personal network into clusters of personal nodes that are formed and are dynamically updated, consisting of core and master nodes.

Recently standardization bodies have started to study the incorporation of self-organization capabilities, like the 3GPP, in Long Term Evolution (LTE) and System Architecture Evolution (SAE) Operations, Administration, and Maintenance (OAM) mainly for Home NodeBs and eNodeBs [71].

2.4.1 Self-Organizing Networks in 3rd Generation Partnership Program

In the networks that 3GPP standardizes networks the concept of the SON is introduced in order to reduce the OPEX associated with the management of this larger number of nodes from more than one vendor [8]. Automation of some network planning, configuration and optimization processes via the use of SON functions can help the network operator to reduce OPEX by reducing manual involvement in such tasks. Different architectures that are possible for implementing various SON use cases have been identified:

- Centralized SON: SON solution where SON algorithms are executed in the OAM system. Centralized SON has two variants:

  - NM-Centralised SON: SON solution where SON algorithms are executed at the Network Management level.
  - EM-Centralised SON: SON solution where SON algorithms are executed at the Element Management level.
- Distributed SON: SON solution where SON algorithms are executed at the Network Element level.
- Hybrid SON: SON solution where SON algorithms are executed at two or more of the following levels: NE or EM or NM.

The vision is that algorithms automate tasks that currently require significant planning efforts. In parallel, the 3GPP works on specifications for 3G LTE, and SON is central in the network management and optimization discussions. SON can mean vastly different things, but three components are central [72]:

- **Self-configuration**: plug and play functionality where network elements are configured (identity allocation, software upgrade, communication link establishment, etc) automatically.
- **Self-optimization**: more or less continuous adaptation of parameters to meet specified requirements typically specified at a high level.
- **Self-healing**: algorithms to handle disruptive events and to minimize negative consequences on services.

Figure 2-4 presents the SON use cases which are foreseen by operators in the context of NGMN [73].

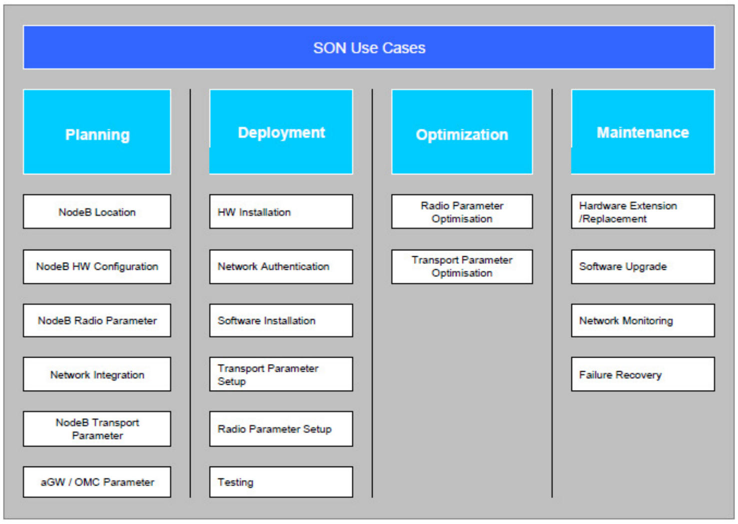


Figure 2-4: NGMN: Categories and sub-groups of SON related use cases [73]

2.4.2 Self-Organizing Networks in Institute of Electrical and Electronics Engineers

The active IEEE 802.16m task group [74], has introduced a self-organizing functionality for radio resource control and management so as to automate the configuration of base station or relay station parameters and to optimize the network performance, coverage, and capacity. Furthermore, in the context of self-organizing and auto-configuring mesh Networks, IEEE 802.11s [75], which is an IEEE 802.11 amendment for mesh networking, defines how wireless devices can interconnect to create a WLAN mesh network, which may be used for static topologies and ad-hoc networks.

2.5 Wireless Network Resources Self-Optimization

Wireless Resource Management is a fundamental aspect of any wireless communication system, which could be addressed at two different levels:

a. a priori allocation of resources to the network nodes (e.g., the split of the resources (e.g., frequencies, slots) among the different available radio access technologies in a classical network planning process).

b. (re-)assignment of the available (previously allocated) resources according to the QoS needs.

Regarding the functionality in (a), up to now it has been carried allocating in a fixed way the estimated required resources, with limited possibility of modifications on a short/medium term basis. After some period of time, network is manually re-planned depending on traffic variance. Point (b) refers to what is usually performed within the Radio Resource Management (RRM) functionalities of the network nodes. Below you can see a non exhaustive list of the self-optimization use cases that 3GPP addresses for LTE networks [76]. The majority of them are also applied for other types of networks (WiMAX, WiFi):

- **Coverage and capacity optimization**: A typical operational task is to optimize the network according to coverage and capacity. Planning tools support this task based on theoretical models but for both problems measurements must be derived in the network. Call drop rates give a first indication for areas with insufficient coverage, traffic counters identify capacity problems.

- **Energy Savings:** A typical critical cost for the operator is the energy expenses. Cuts on energy expenses could be realized if the capacity offered by the network would match the needed traffic demand at any point of time as close as possible.

- **Interference Reduction:** Capacity could be improved through interference reduction by switching off those cells which are not needed for traffic at some point of time, in particular home eNodeBs when the user is not at home.
- **Mobility robustness optimization:** Manual setting of Handover Optimization (HO) parameters in current 2G/3G systems is a time consuming task. In many cases, it is considered too costly to update the mobility parameters after the initial deployment. For some cases, RRM in one Evolved Node B (eNB) can detect problems and adjust the mobility parameters, but there are also examples where RRM in one eNB cannot resolve problems. Incorrect HO parameter settings can negatively affect user experience and wasted network resources by causing HO ping-pongs, HO failures and radio link failures (RLF). While HO failures that do not lead to RLFs are often recoverable and invisible to the user, RLFs caused by incorrect HO parameter settings have a combined impact on user experience and network resources. Therefore, the main objective of mobility robustness optimization should be reducing the number of HO-related radio link failures. Furthermore, non-optimal configuration of handover parameters, even if it does not result in RLFs, may lead to serious degradation of the service performance. Example of such a situation is incorrect setting of the HO hysteresis, which may be the reason for either ping-pong effect or prolonged connection to non-optimal cell. Thus the secondary objective will be reduction of the inefficient use of network resources due to unnecessary or missed handovers. HO-related failures can be categorized as follows: a) Failures due to too late HO triggering, b) Failures due to too early HO triggering, c) Failures due to HO to a wrong cell. Additionally cell-reselection parameters not aligned with HO parameters may result in unwanted handovers subsequent to connection setup, which should be avoided by parameter adjustments done by MRO function.
- **Mobility Load balancing optimization:** Optimization of cell reselection/handover parameters in order to cope with the unequal traffic load and to minimize the number of handovers and redirections needed to achieve the load balancing. Self-optimization of the intra-LTE and inter-RAT mobility parameters to the current load in the cell and in the adjacent cells can improve the system capacity compared to static/non-optimized cell reselection/handover parameters. Such optimization can also minimize human intervention in the network management and optimization tasks. The load balancing shall not affect the user QoS negatively beyond what a

user would experience at normal mobility without load-balancing. Service capabilities of Radio Access Technologies (RATs) must be taken into account, and solutions should take into account network deployments with overlay of high-capacity and low-capacity layers where high-capacity layer can have spotty coverage. Load balancing can be done either in Intra-LTE load balancing or Inter-RAT load balancing.

- **Inter-cell Interference Coordination:** In reuse one cellular networks mutual interference between cells occurs. Within the OFDM and SC-FDMA based LTE system interference has to be coordinated on the basis of the physical resource blocks (PRBs). Such interference can be reduced or avoided in uplink and downlink by a coordinated usage of the available resources (PRBs) in the related cells which leads to improved SIR and corresponding throughput. This coordination is realized by restriction and preference for the resource usage in the different cells. This can be achieved by means of (ICIC) related RRM mechanisms employing signaling of e.g., HII, OI and downlink (DL) TX Power indicator. ICIC RRM might be configured by ICIC related configuration parameters like reporting thresholds/periods and preferred/prioritized resources. Then these have to be set by the operator for each cell. Setting and updating these parameters automatically is the task of a SON mechanism.
- **RACH Optimization:** The Random Access CHannel (RACH) configuration has critical impacts to system performance. The RACH collision probability is significantly affected by the RACH settings, making this a critical factor for call setup delays; data resuming delays from the uplink (UL) unsynchronized state, and handover delays. It also affects the call setup success rate and handover success rate. Since UL resource units need to be reserved exclusively for RACH, the amount of reserved resources has impacts on the system capacity. A poorly configured RACH may also result in low preamble detection probability and limited coverage. Therefore, RACH parameter optimization provides significant benefits to the deployed network. The setting of RACH parameters depends on a multitude of factors, (e.g., the uplink inter-cell interference from the Physical Uplink Shared Channel (PUSCH), RACH load (call arrival rate, HO rate, tracking area update, traffic pattern and population under the cell coverage as it affects the UL synchronization states and hence the need to use random access), PUSCH load, UL and DL imbalances. Since these are affected by network configuration (e.g.,

antenna tilting, transmission power settings and handover thresholds), any change in these configurations would also affect the optimum RACH configuration. For example, if the antenna tilting of a cell is changed, the coverage of cells in the vicinity will be changed, consequently affecting the call arrival rate and handover rate at each cell. This will affect the amount of RACHs in each cell, including the usage per range of preambles. Then, the operator will have to check the RACH performance/usage in each cell and detect any problems on RACH associated with the applied changes. If required, it may further trigger some adjustments in RACH configuration. Measurements on the RACH performance/usage are needed to be collected at a SON entity. An automatic RACH optimization function monitors the prevailing conditions, e.g., a change on RACH load, uplink interference, and determines and updates the appropriate parameters.

2.5.1 Coverage and Capacity Optimization, and Energy Saving

Energy efficiency has been extensively studied in specific research areas, e.g., wireless sensor networks, where sensor battery life is a crucial parameter for WSN duration [77]. However, in recent years, energy efficiency has been a fundamental concern for almost every ICT field e.g., data centers [78], wired networks [79], wireless networks [5]. The reduction of energy consumption in a wireless communication system is a complex exercise because of the various tradeoffs and interdependences [80]. For that reason, recent works attempt to elaborate on the various levels of energy consumption, by covering network level aspects, including deployment, architecture and network management; link level perspective; and the component level, including hardware implementation [81]. Chen et al. in [5] and Han et al. in [82] highlight the role of the base station in the total power consumption of a wireless network and specifically the high energy need of the power amplifier. In the context of EARTH project [83], Auer et al. [84] quantify the power consumption of mobile communication systems, by discussing power models that map the radiated Radio Frequency (RF) power to the supply power of a BS site, taking into account various traffic models (low, medium, high load, peak traffic demand) and deployment cases (e.g., urban, suburban, rural). Also, a holistic system view design is described by Strinati et al. [85] in order to ensure that any proposed solution to improve energy efficiency does not degrade the energy efficiency or performance on any other part of the system. The authors are focusing on technical solutions related to resource allocation strategies designed for increasing diversity order, robustness and effectiveness of a wireless multi-user communication system.

Various methods for the reduction of energy consumption in wireless communication systems that are related to CCO have been proposed recently. Samdanis et al. [86] presents a centralized and a distributed algorithm for energy saving in radio access networks, based on the concept of energy partitions, which are associations of powered on and off BSs formed by a collective, decision of network element. The objective is to match the overall offered bandwidth in terms of coverage and capacity in a dynamic manner. The simulation results suggest that generally centralized algorithms perform better especially as bandwidth demand per user increases because it is easier and more effective to shift active sessions among cells at once. Also, in [84] the problem of minimizing the power consumption of wireless access networks by switching on and off and adjusting the emitted power of access stations based on different traffic profiles that can be experienced by the network, is addressed via an integer linear programming model that allows to select consumption and to guarantee coverage and enough capacity to serve active users in the service area. Moreover, in [88] the authors propose deployment strategies on the power consumption of mobile radio by exploiting micro base stations per cell in addition to conventional macro sites; the metric of area power consumption as a system performance metric is introduced. In a full traffic load scenarios, the use of micro base stations has a rather moderate effect on the area power consumption of a cellular network. Cao et al. in [89] consider multi-BS cooperation and wireless relaying technologies and specifically how the energy saving performance is affected by the system parameters: traffic intensity and network density.

Jardosh et al. [87] propose resource on-demand (RoD) WLAN strategies that can efficiently reduce energy consumption of a WLAN without adversely impacting the performance of clients in the network. Authors conjecture that energy-efficient mechanism for large-scale and high-density WLANs should be designed and developed today – to save energy in future WLANs and thus avoid the escalation of energy wastage. Optimized downlink communications in order to reduce energy consumption at the base station is investigated in [87]. Authors support that cooperation between end users and network operators may result in significant overall energy savings. An auction based pricing system that incentivizes both parties to cooperate for energy savings has been adopted.

UE energy efficiency in conjunction with the network-side energy saving techniques is necessary requirement for building the whole network energy view, and an area where further research work could take place. Taking into account the above analysis of the literature work it is obvious that important effort needs to be dedicated to the understanding of energy consumption problems, moving from single network elements (e.g., base station energy consumption) towards a more holistic network level approach.

2.5.2 Dependencies between Self-Optimization Processes in Self-Organizing Networks

In the context of a SON the performance metrics of the various configuration or optimization problems and the associated configuration actions have various interactions and dependencies. A conflict in a control parameter or configuration action (e.g., AP switch On/Off, antenna tilt) appears if two or more metrics, which exceed a predefined threshold, attempt to change the same control parameter towards different (i.e., conflicting) directions. On other hand, a conflict on a performance metric (e.g., interference, load) indicates that a triggered configuration action has a negative influence on another performance metric. This might lead to a sequence of adaptations creating a ripple effect situation. In [90] Gruber et al. provide the following definitions for parameters value conflict and metric value conflicts:

- Parameter value conflicts occur if two use cases have access to the same control parameter.
- Metric value conflicts occur if any two use cases influence a common metric that is used as feedback information to influence either use case.

Figure 2-5 illustrates how the authors in [93] have visualized the control parameter conflict and conflict due to observability dependency. The latter term is used by authors to describe the metric value conflicts. In addition, authors in [93] identify two cases for the control parameter conflict: a) directionally conflict, where e.g., SON function A wants parameter A to increase, while SON function B wants parameter A to decrease, b) magnitude conflict, where SON function A wants a change of parameter A that has a different intense (larger, smaller) comparing to SON function B.

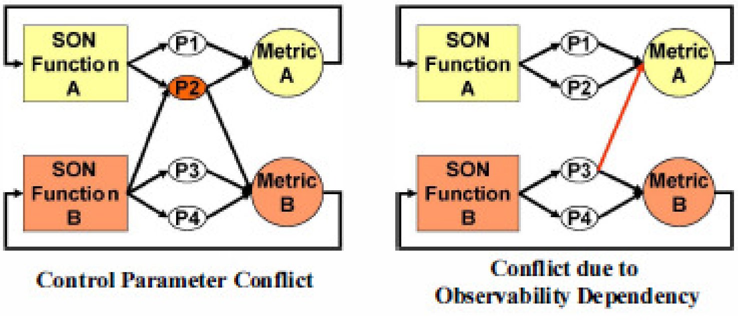


Figure 2-5: Conflicts types [93]

Up to now, the work on SON has focused mainly on the development of individual or stand-alone functionalities. Existing network management systems do not provide solutions for the interaction of multiple SON functions. On the other hand, a lot of research work has been performed for the joint optimization of specific problems e.g., [94]. In addition, policy-based systems have been studied for the detection and the resolution of conflicts. Initially, authors in [95] addressed policy conflicts in policy-based management as being analogous to software bugs. Charalambides et al. [96] propose the use of Event Calculus and formal reasoning for the analysis of both static and dynamic conflicts in a semi-automated manner, which is applied in the domain of DiffServ QoS management. In [97] policy conflict analysis entails analyzing a candidate newly created or modified policy on a pair-wise basis with already deployed policies and potential conflicts between the policies are fed back to the policy author. Central to the approach is a two-phase algorithm which firstly determines the relationships between the pair of policies and secondly applies an application specific conflict pattern to determine if the policies should be flagged as potentially conflicting.

For efficient coordination of SONs optimization or configuration problems it is important to identify the various interactions between control parameters (i.e., configuration actions) and performance metrics. Towards this goal, in UniverSelf FP7 project [98] Figure 2-6 has been designed [99], illustrating those interactions and conflicts of the LTE SON use cases that are described in [76]. The use cases themselves are depicted in yellow, the potentially affected control parameters are depicted in pink, whereas the influenced metrics are shown in light blue.

The avoidance of conflicts either on the parameters (i.e., adaptations, re-configurations) or on the metrics and the joint assessment of the various SON problems is a challenging issue for SONs. Different methods have been recently proposed in order to coordinate conflicts between SON use cases. In [100] it was proposed to jointly optimise tightly coupled use cases and to otherwise use a trigger strategy, i.e., a use case completes the optimisation and then triggers a use case(s) to re-optimise. In contrast to this approach, in [90] a separation strategy is proposed, according to which use cases are grouped in time domains. Within the time domains a joint optimisation would be performed; slower processes would have priority over faster processes as faster processes are more flexible to react. A SON Coordinator framework is proposed in [93] with different functional roles that aim at achieving the goal of harmonized SON system operation, according to the high-level performance objectives of the operator. The

SON coordinator intends to ensure that the individual SON functions jointly work towards the same goal, formulated by the operator's high-level objectives. This can be achieved by effectively and appropriately harmonizing policies and control actions of the SON functions. In [101] the authors focus on SON interactions, aiming to provide a framework for understanding and analyzing them. Since the RAN ecosystem is increasingly heterogeneous and multi-technology with the introduction of new Radio Access Technologies (RATs), a heterogeneous/multi-technology setting has been adopted with two basic and important optimization tasks: Admission Control (AC) and RAT Selection (RS). A basic understanding of the interaction between these two SON mechanisms will be provided with an initial assessment of a simple interaction control.

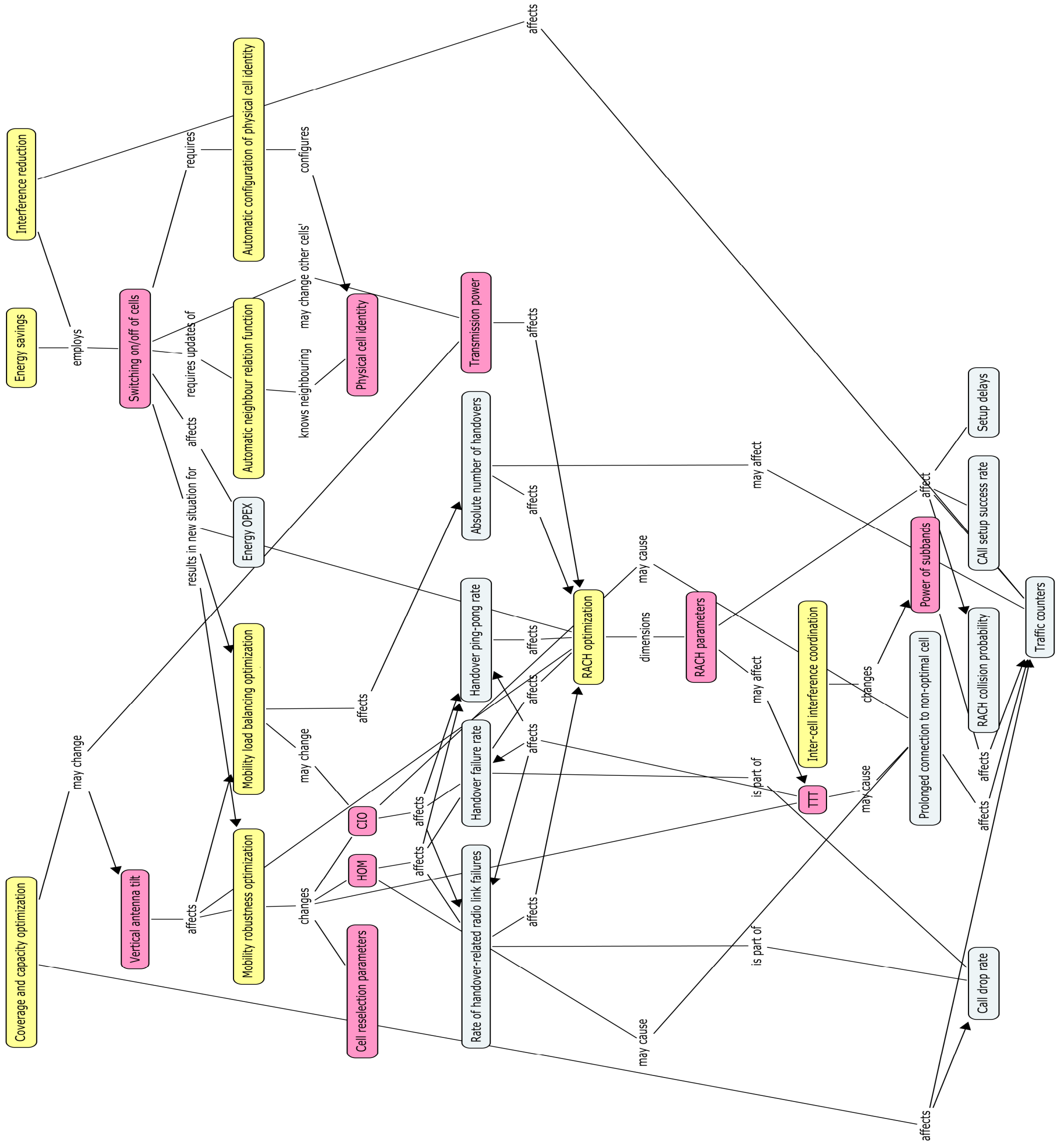


Figure 2-6: Interaction and conflicts of 3GPP SON use cases [99]

CHAPTER 3

SYSTEM ARCHITECTURE FOR A COGNITIVE AND SELF-ORGANIZING NETWORK

This section defines the functional and software architecture for the development of a self-organizing cognitive communication system. Specifically we introduce the cognitive network element cognitive manager (level I) and the network domain cognitive manager (level II), which functionalities are analyzed. Both entities are agents that instantiate the main features of the cognitive cycle for a communication network system. The deployment and assessment of the framework on a WiFi/WiMAX wireless environment are presented. Finally, results and the lessons learned from the experimentation activities are discussed and recommendations are proposed.

3.1 Principles

Cognition is central to Future Internet self-managed systems. An in-network cognitive cycle allows communication systems to dynamically adapt their network parameters (e.g., queue sizes, radio frequencies) according to the varying network conditions (e.g., changes in topology, traffic demands), and to improve their reasoning capabilities by exploiting the feedback from previous events or from historic data that are stored locally [102].

Embedding cognition in the network appliances requires the design of cooperative closed control loops that will constantly monitor the network state, exchange information with their neighborhood to build a consistent view of the ongoing network situation, select reconfiguration actions and acquire more elaborated knowledge through learning. It has been established from the past research activities that a cognitive control loop or cycle shall be decomposed into two parts, each operating at different time scale (Figure 3-1):

- A reactive part that senses and acts directly on the network device based on predefined rules and according to network administrators policies (shorter time scale).
- The learning part that exploits feedback from previous events or historical data to extract knowledge and adapt accordingly the decision rules (longer time scale).

3.2 Hierarchical Distribution of the Cognitive Cycle for Network Systems Management

Future network systems design principles are based on high autonomy of network elements in order to allow distributed management, fast decisions, and continuous local optimization. The generic Cognitive Cycle model is envisaged to be in the heart of Future

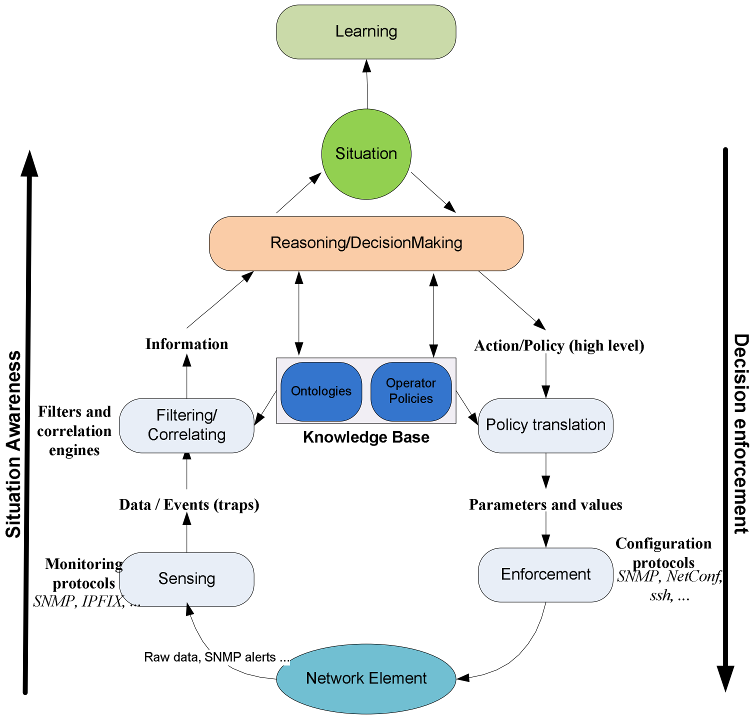


Figure 3-1: Cognitive cycle/Control loop

Internet Elements (e.g., access points, base stations) and it leads to their autonomy [103].

The cognitive cycle can have various levels of realization from simple autonomic cycle (simple, reflex-based Monitoring-Decision Making-Execution) to a fully realized cognitive cycle (Monitoring-Situation awareness-Reasoning-Decision Making-Execution-Learning). The computational capabilities and the specific networking problems that are addressed by individual future Internet elements are some of the criteria for the selection of the appropriate realization level.

For the specification of a holistic self-management architectural framework, the Distributed Cognitive cycle for System & Network Management (DCSNM) is proposed (Figure 3-2). DCSNM follows a hierarchical distribution of cognitive cycles, breaking down the respective functional entities and mechanisms for solving network management problems and other self-management operations (e.g., learning, monitoring) to:

a) network element (e.g., access points, routers),

b) network cluster, which is communities of network elements,

c) network domain, which is structured and long-term federation of network elements,

d) network management system that controls the underlying entities and provides the human-to-network interface for the administrator.

Apart from the vertical interactions among the above mentioned layers, interactions are taking place (horizontally) among peer entities of the same layer (e.g., network elements cooperation, network domain managers) that may be owned by different network operators. This hierarchical approach is considered as one of the key principles to handle complexity and for effective localized and scalable management. Hierarchy is a fundamental characteristic of natural systems, which exhibit high degree of organization.

The logic behind the introduction of the DCSNM is to serve as the conceptual template of introducing the Future Internet mechanisms advances in the overall system as well as a network management instrument. Hence, it is a formulated tool for addressing the complexity and capabilities of networks, services and management elements and their roles as providers of new paradigms that are emerging in the evolution of needs and mechanisms for Future Internet, service and network infrastructures in general. The DCSNM can be used as the guiding framework for constricting the architectural and functional features in relevant deployment scenarios.

The decomposition of network management into responsibility areas provides the means for the development of a management architecture which goal is the reduction of human interventions in the fundamental management functions. These include mechanisms that render the networks capable to configure, optimize, heal and protect

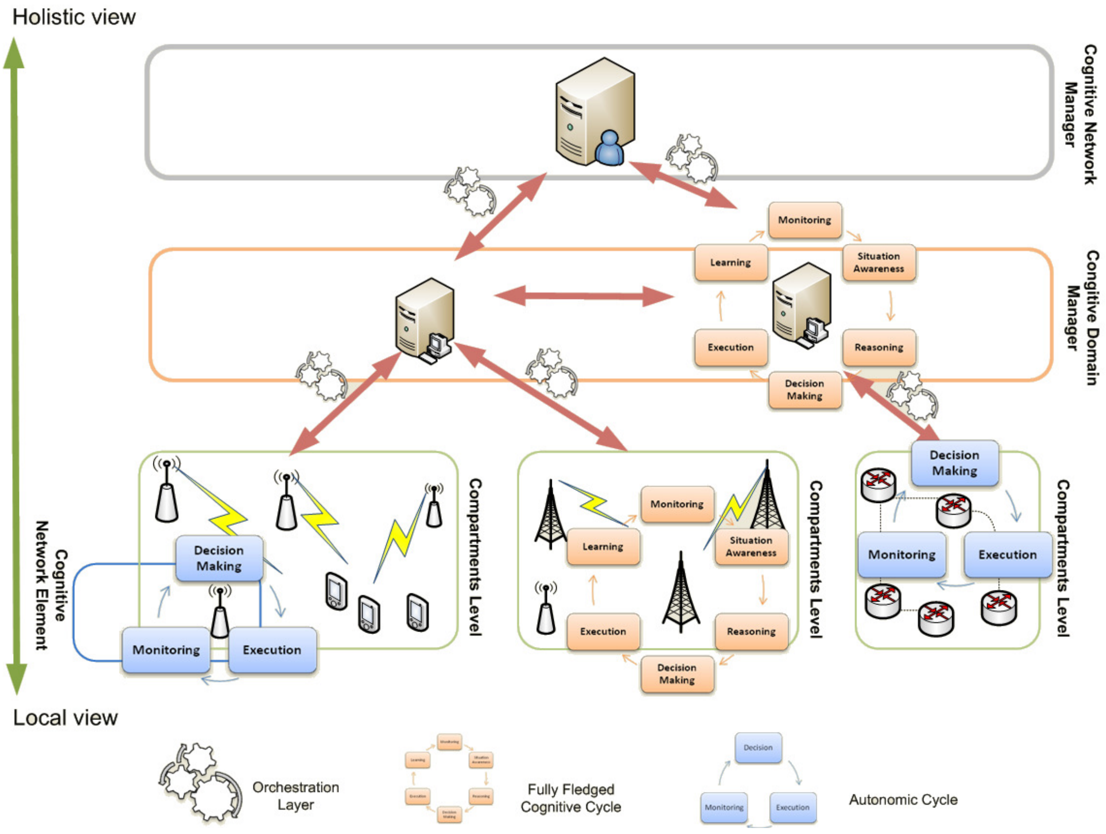


Figure 3-2: Distributed cognitive cycle for system and network management

itself, but also handle the emerging complexity. Such a decomposition combined with the introduction of cognitive functionalities at all layers will allow decisions and configuration at shorter time-scales. Each element at the identified layers has embedded cognitive cycle functionalities and also the ability to manage itself and make local decisions. For an efficient and scalable network management, where various stakeholders participate, a distributed approach is adopted. Dynamic network (re)-configuration is based on the cooperation of different network elements and or even service components. Hints and requests/recommendations are exchanged among the layers, in order to indicate a new situation or an action for execution.

3.3 Functional Architecture

The cognitive cycle (feedback loop) can have various levels of realization according to the type of the device and the hierarchical level that it is placed. Each cognitive cycle has the ability to expand and consider the results of other cognitive cycles in a collaborative manner, thus leading to an incremental development of local (element level) and global knowledge (network wide). Figure 3-3 shows some of the main elements of the cognitive cycle focusing on steps required for reaching a decision on taking a specific execution action by the element. One of the main prerequisites for achieving the steps of the cognitive cycle is in the self-awareness plane, seen as the product of the continuous activity in network elements, and their interactions with neighboring elements. From the practical point of view and its representation in the cognitive cycle, self-awareness is seen as the elements' view on the internal and external processes, statuses and states [104], [105], [106], [107].

Monitoring feature is intended to provide the basis for intelligent deduction processes in network elements represented in the situation awareness. Situation awareness is the additional ability to know and deduce what is happening in the network (e.g., faults, optimization opportunities), involving the comprehensive set of data inputs and related to the environment in consideration. Situation awareness is the step that precedes and constitutes the foundation for decision-making. Its ultimate result is a validated situation being the point at which an instance of situation awareness process is completed hence progressing to the decision-making part. Thus, decision making has the full awareness of

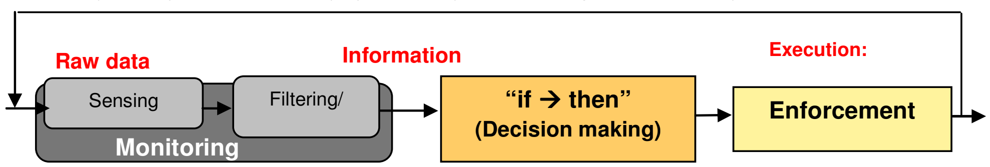


Figure 3-3: A simplified representation of the M-D-E cycle in the basic automated "reflex" case

the system statuses represented in the situation resembling a view or comprehension that a network administrator would have when a certain event (trigger)/information/data is obtained and analyzed from the current operational aspects in the system.

Monitoring leads to Level 1 of situation awareness, which is the stage for Characterization of Operational States/Perception (e.g., high interference, high loaded network area). Next level, Level 2 of situation awareness realizes the assessment of the environment, analogous to network administrator response to an event/information. When there is a situation trigger the environment needs to be assessed in order to prepare for decision-making. Level 3 of situation awareness is called Projections and includes prediction on what would happen in the future based on the current assessment of the environment and/or what conditions should be met to proceed to decision-making. These steps precede decision making where actions are specified subject to the specific invocation of the cognitive cycle.

The steps described above use self-awareness as the real-time snapshot of system states and in addition consult the knowledge base instructions, as depicted in Figure 3-4. The knowledge base can contain instruction for situation awareness level 1 - Interpretation Library of Operational States ultimately generating a Situation Trigger. For situation awareness level 2, knowledge base contains Procedures for Assessment of the Environment, i.e., analogous to network administrator training. A similar mapping of knowledge base usage can be observed for other elements of the cognitive cycle e.g., decision making. Figure 3-5 depicts the instantiation of the logical architecture on a wireless access point.

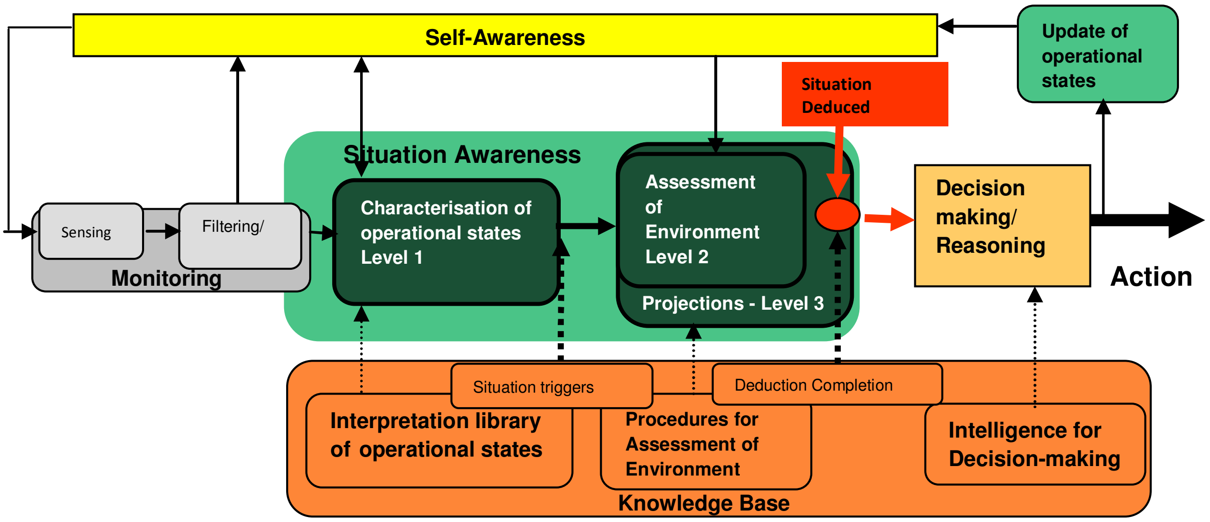


Figure 3-4: Cognitive cycle logical architecture

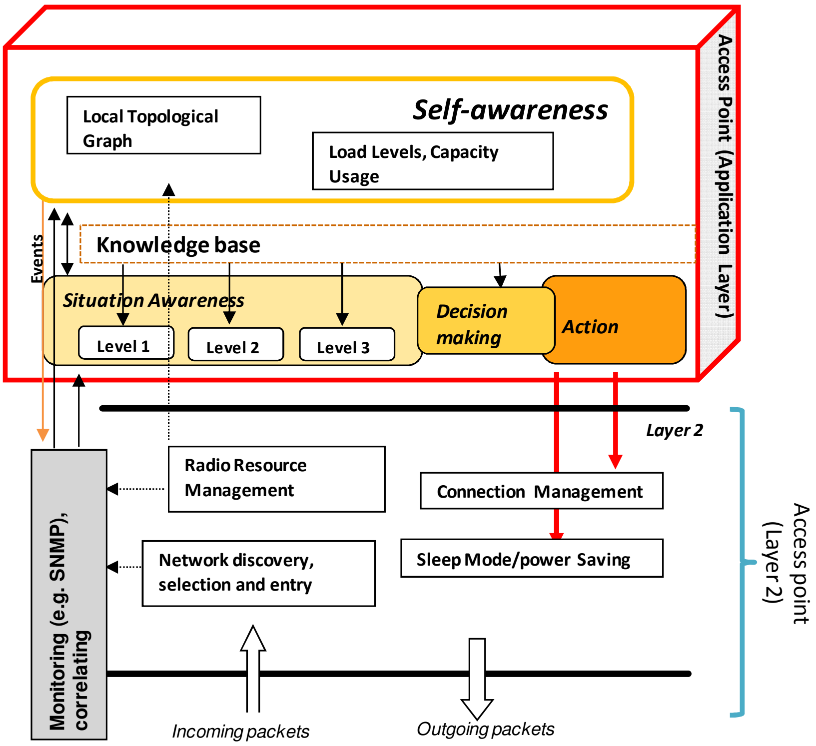


Figure 3-5: Instantiation of an access point control loop for coverage optimization

The software module that encompasses the cognitive control loop functions is further referred to as the CNM. The CNM can detect device anomalies or network service disruptions, diagnose root causes, and compute possible corrective actions to finally select and enforce those actions that best fit the system or operator’s objectives. Cognitive network elements should be able to represent relationships among objects/events in order to support inference and/or reasoning tasks (Figure 3-6). Specifically, each cognitive manager should keep locally the associations between the available metrics (MIBs, measurements), provided by the monitoring tools of a network node, and the problems that have been identified for the respective node. Moreover, the correlations between the metrics and the performance goals that the network node has are also kept locally, as well as the configuration actions that will provide the remedy in the case of an identified fault or for the achievement of an optimization opportunity, as the latter is described by the specified goals. This knowledge base is an attempt to transfer the awareness that a human network operator has developed through experience and training and particularly the rules that instantiate the various correlations. The knowledge base is adapted according to the device type (base station, router etc), and the management problems that are addressed.

Network devices are physically distributed and the configuration of a particular element often interplays with the configurations of other nodes. As a consequence, the network

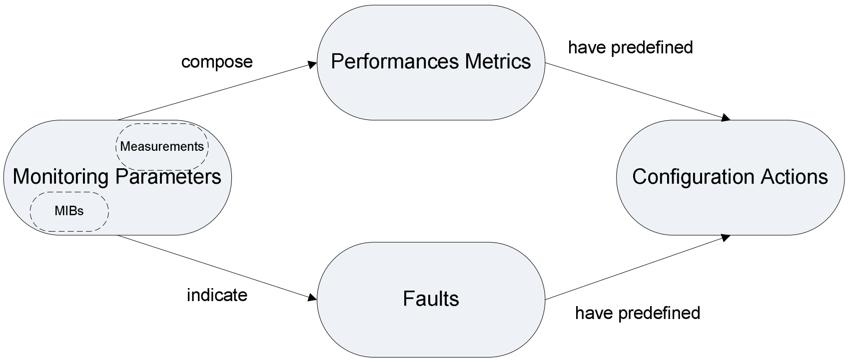


Figure 3-6: Knowledge base generic structure

context that each CNM uses for making decisions cannot be limited to its local point of view. A centralized approach to provide this broader picture would not be effective, for scalability and fault tolerance reasons. Thus, local cooperation among neighboring self-managed elements is required so as to exchange information that is necessary for the decision making process (Figure 3-2).

If broader information is needed for making some decisions (e.g., regarding global traffic load), a Domain CNM takes over. The Domain CNM is a more sophisticated CNM which manipulates higher level knowledge. The domain knowledge contains monitored metrics collected by a set of subsuming CNMs, which are processed and analyzed. It also includes policy rules that describe the relations among problems and associated potential solutions of the target domain. Having this richer knowledge, potentially complemented with interactions with other Domain CNMs, it can proceed with global decision making, thus solving problems that standard CNMs cannot address locally. Domain CNMs adjust policies to fit the desired system behavior and learn from previous experiences to improve the impact of future decisions.

A two-tier architecture, both distributed and hierarchical, has been introduced for broadband wireless networks [53]. In this context two types of CNMs have been identified:

- simple CNM that is referred to as NECM,
- domain CNM entitled NDCM.

NECM implements the cognitive cycle at the network element level, providing an intelligent adaptation layer to the conventional control plane. Management problems that cannot be addressed directly at the network element level, due to computational or communicational constraints, are escalated to the respective NDCM level. The NDCM incorporates the required cognitive capabilities to identify optimization opportunities and solve problems that require a greater view of network status, the cooperation between neighboring domains, or even the resolution of conflicts in metrics/parameters. The NDCM does not disturb the distributed and hierarchical nature of the proposed architecture; it is not a centralized physical entity. Various NDCMs could exist in a specific network domain, hosted even at the available network devices (e.g., router, access point). There are various criteria for the selection of the number of NDCMs e.g., the type of the network technology, the purpose of the network element or other spatial and geographical features.

Each NECM periodically retrieves cross-layer monitoring data by the monitoring tools that each network element supports, while monitoring data are collected by neighboring network elements (i.e., NECMs). Locally collected monitoring data are also periodically transmitted to the NDCM. Each NECM periodically triggers the process that undertakes to update its self-awareness. This phase, entitled “Inference for Faults and Optimization Opportunities Identification” includes the interpretation and the correlation of various network parameters, identifying faults or performance optimization opportunities. The outcome is a list of performance metric with a calculated value, which indicates the degree up to which a network element faces a specific situation (e.g., high load, low interference). This phase provides the input for the second inference phase (i.e., Configuration action selection), where according to the calculated values of the

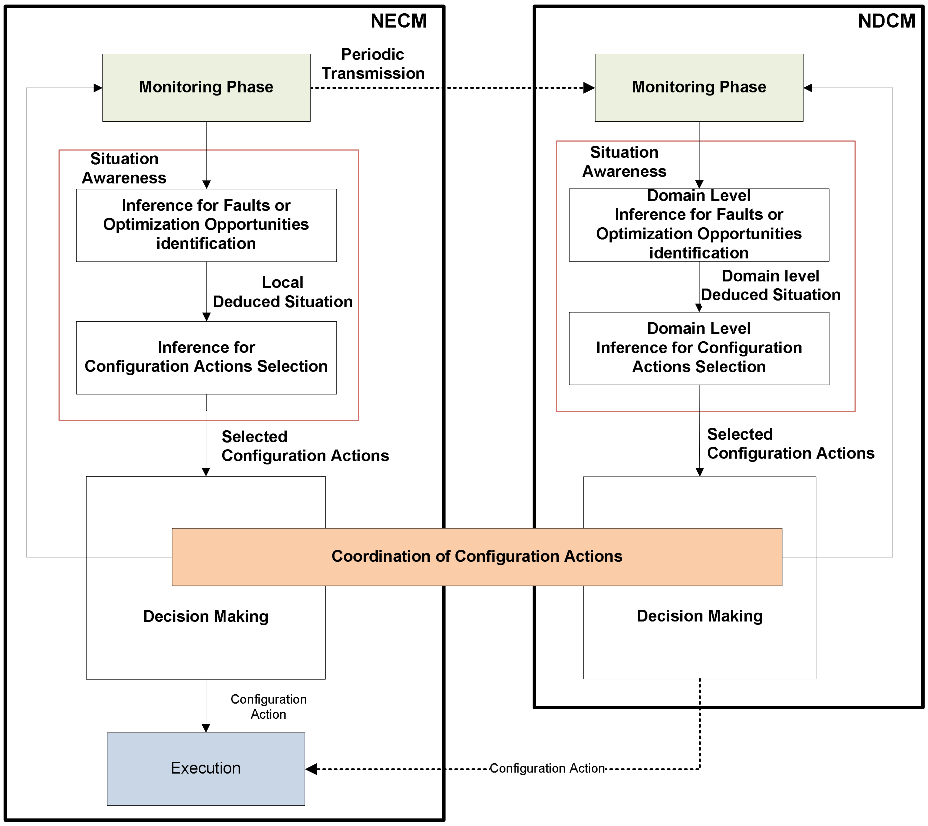


Figure 3-7: Network Element and Domain Cognitive Manager activity diagram and interactions

performance metrics and the set policy rules the appropriate configuration actions are selected. In the case that more than one configuration actions are triggered in the same or neighboring device, then the functionality for the coordination of SON problems undertakes to resolve possible conflicts or dependencies, by prioritizing the most appropriate action. The NECM transfers the selected type of configuration action to the execution mechanisms of the corresponding node and applies the decided adaptation.

Moreover at the NDCM level, after the pre-processing and filtering of the monitoring data that are collected by the NECMs, the domain level “Inference for Faults and Optimization Opportunities Identification” is periodically triggered in order to identify faults or optimization opportunities that local NECM cannot deduce due to limited network view. The NDCM based on the information that has been collected by several NECMs perceives the status of a particular network area. Similarly to the NECM case the NDCM identifies the appropriate configuration action, resolving possible conflicts or other dependencies and sends the decided reconfiguration action to the NECMs for the enforcement of the adaptation. Furthermore, NDCM improves the inference and reasoning capabilities of itself as well as of NECMs, by exploiting the feedback from previous events or from historic data that are stored locally.

3.4 Cognitive Network Manager Software Architecture

For the purpose of validating the CNM concept its software architecture has been designed [108]. It carries out the different steps of the cognitive cycle and is conceived to minimize its fingerprint on the embedding devices and to favor reuse and extension. The architecture is modular, enabling the load and replacement of required functions. In addition, the modules are mainly platform and technology agnostic and APIs are carefully designed to reach genericity. The main modules of the CNM software architecture, as illustrated in Figure 3-8, are: the discovery service, the communication service, the topology service, the behavior container, the blackboard, the network element Controller, and the fuzzy logic (FL) engine (Figure 3-8).

All CNMs (NDCMs, NECMs) implement three basic device agnostic services which enable them to discover their peers, build a view of the cognitive hierarchy, and communicate with their neighbors. The Discovery service is in charge of the discovery of the neighborhood physical topology based on a probing approach. This physical network topology is used to identify the network addresses of the neighboring nodes and to evaluate the impact of a link or node failure on the active network services. The Topology service maintains information concerning neighboring CNMs (network addresses,

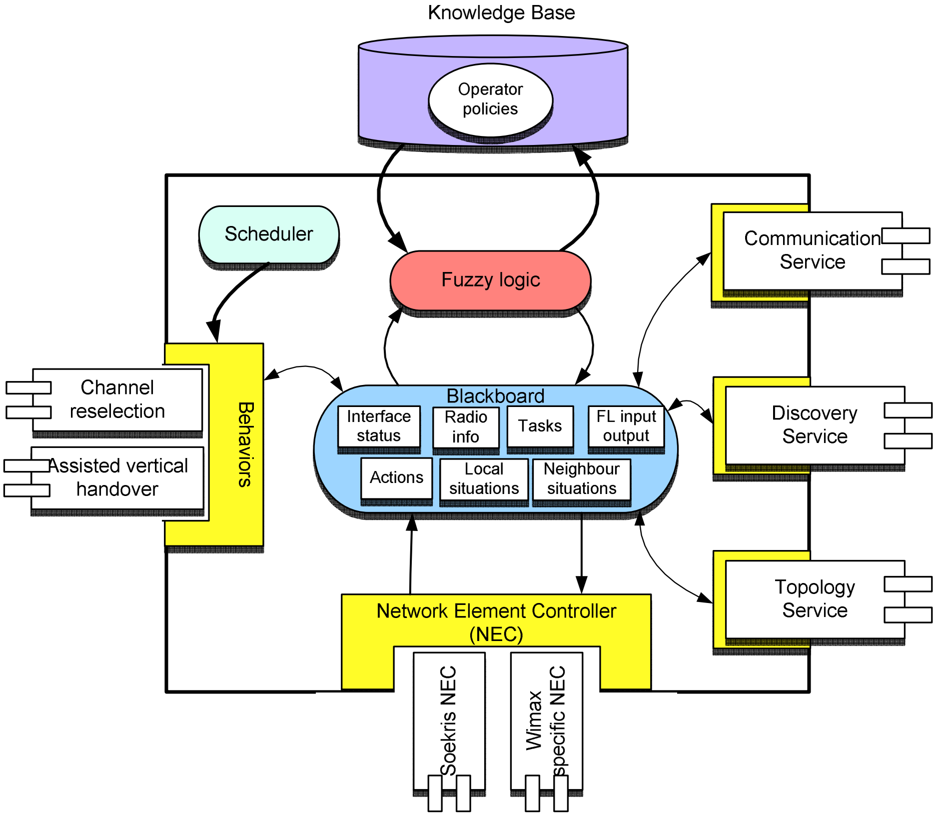


Figure 3-8: Network Element Cognitive Manager

network interfaces used) necessary to support cooperation, coordination and knowledge exchange, after the identification of the physical topology. Finally, the Communication service provides communication facilities to the CNMs, which is in charge of message exchange among NECMs and between NECMs and NDCMs.

The Behavior container hosts behavior modules which perform tasks or computations inherent to specific objectives in order to optimize, configure or heal network elements. It is managed by the Scheduler that triggers the execution of registered schedulable components according to their behavioral profile and the identified network situation. The behaviors can work in three modes: a) one shot mode for tasks that the CNM invokes for urgent events, b) periodic mode for operations that are triggered in predefined time slots, and c) publish/subscribe mode for the update of the blackboard topics.

The Blackboard is used to provide a shared memory and acts as the knowledge bus of the CNM by providing writing and reading facilities to the different modules. It is organized as a set of topics, where information is published for the subscribed modules to read. They can also partially or totally delete information contained in topics they have registered to. NECs publish the information they produce in the Interface status and Radio information topics for the monitored parameters of supported radio access technologies. The services publish information on the Local situations and the Neighbor situations topics for the status (i.e., problems, events) of the local and neighboring CNMs, respectively. These topics are mainly followed by the behaviors and the FL engine. In addition, the behaviors retrieve from the FL input/output topic the information provided by the FL engine and expose the result of their tasks or actions in the topics with these two labels. Actions topic keeps the list of the configuration actions that are supported by the network device, while Tasks topic stores the tasks that are registered with the scheduler of the CNM.

A FL engine is used for reasoning and decision making. Changes in the set of monitored parameters are published as dedicated topics in the Blackboard. They are analyzed by the FL engine which then publishes its assessment of the device situation on the Blackboard, eventually triggering behaviors and turning the decisions into a sequence of atomic corrective actions.

The Network Element Controller (NEC) manages device-specific sensors and actuators and is only to be found in NECMs. Sensors are in charge of collecting monitored parameters such as the allocated channel in an access point. Actuators are capable of enforcing reconfigurations in the network device. The collected data, adopting a model-based translation scheme, are semantically annotated by the NEC, and published in the Blackboard. Actions are retrieved by the actuators from the Blackboard.

The architecture of an NDCM is similar to the one of an NECM. NDCM manipulates richer knowledge, resulting from data provided by NECMs or coming from exchanges with other NDCMs. Computations and decisions an NDCM makes are different from those an NECM processes. Being at a higher level, the response time of their decisions can also be longer.

3.5 Deployment

3.5.1 Deployment Environment Description

In order to experiment with the cognitive management functionalities, we deployed an heterogeneous network environment comprising two radio access technologies, namely IEEE 802.11 and IEEE 802.16.

Soekris devices, which are low-power, low-cost, Linux-based communication computers (433 Mhz AMD Geode LX, 250 Mbyte DDR-SDRAM), act as re-programmable WiFi APs by using IEEE 802.11b/g radio access technology. Four soekris devices have been placed in a working office environment forming a complete network graph, due to the overlaps among the APs coverage area. The network elements are in fixed and pre-

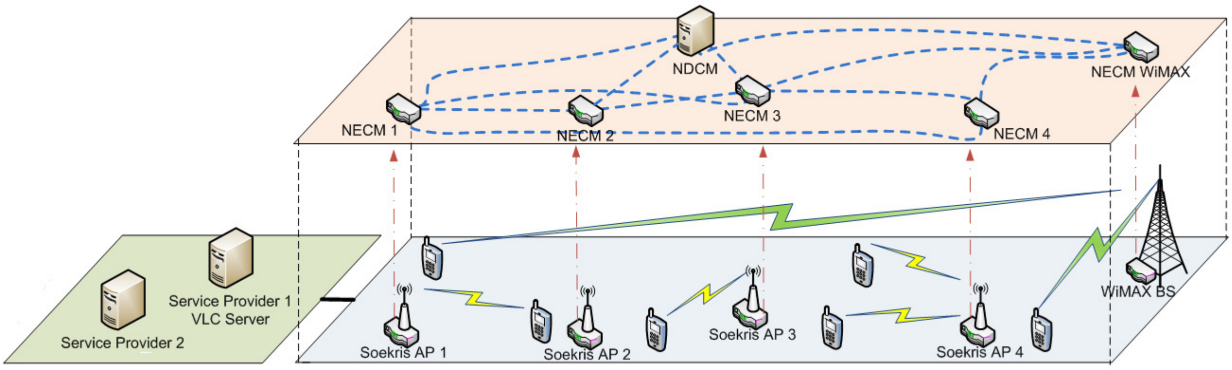


Figure 3-9: Heterogeneous broadband wireless network environment

specified positions with their Tx power set to 25dB. Furthermore, an IEEE 802.16-2004 compliant base station (BS), which operates at the 3.5GHz frequency, is always available; thus covering the whole building, where the network infrastructure has been deployed, with its cell range of 5 kilometers (Figure 3-9).

Forty four single and dual-radio access technology terminals are available in the network area and have Internet connectivity either via WiFi or WiMAX technology. The overall topology is interconnected to a 10/100 Mbit/s backbone network providing Internet connectivity to the associated clients. In addition, the users can consume a video service, provided by a VLC-based multimedia server that resides in the deployed testing facilities and connected to the backbone network.

The upper layer of Figure 3-9 presents the logical graph of CNMs as well as their communication paths that facilitate the local network management. Each network node, either WiFi AP or WiMAX BS, has incorporated an NECM. Specifically, in the case of the WiMAX BS we had to connect an external device, where the WiMAX BS NECM is placed, and have it communicate with the WiMAX BS for monitoring or execution purposes via SNMP GET and SET commands, respectively. An NDCM entity is also deployed in the testing facilities, as a standalone device, which undertakes to communicate with NECMs; acting also as the management bridge between the two different network environments (WiFi, WiMAX).

We have implemented the CNM using OSGi, a Java platform providing functionalities for service-oriented programming. This framework enables to easily manage small modules called bundles and reaches design requirements of modularity and extensibility. Our implementation of the CNM comprises 15 bundles, reflecting the containers and their instances described in Figure 3-8. The NEC bundle interacts with Linux commands. In addition, the ontology is written with the Web Ontology Language (OWL) and the inference rules with the Semantic Web Rule Language (SWRL). Finally, Jess, a well-known Java-based rule engine, is integrated in the Inference bundle.

3.5.2 Test Case

We deploy the cognitive management architecture in the environment described in Figure 3-9. We have considered two management issues from the family of dynamic coverage and capacity optimization. The first one tackles coverage optimization by dynamically reallocating channels of nodes experiencing high interference. The second issue addresses capacity optimization, where users are assigned to different radio access technologies (WiFi or WiMAX) to balance the load across the entire network. In the first case only the NECMs are involved to optimize channel allocation in case of high interference due to overlapping channel assignment. In the second case the domain view of an NDCM is required to balance traffic load between different radio access technologies forcing a vertical assisted handover. A future network environment, especially in urban areas, will include different access networks deployed in the same geographical area. This unstructured network environment results in dense AP topologies, with high coverage or frequency overlapping. The above in conjunction with users' varying traffic volume and service requirements create optimization opportunities concerning network management. Coverage and capacity optimization and operations such as load balancing and interference reduction help towards the avoidance of network resources' overutilization and underutilization. The latter is also supported via the cooperation among heterogeneous networks.

NECMs constantly monitor local parameters such as PER, channel utilization, number of associated terminals. The first stage of the cognitive cycle consists in building knowledge from such raw data. Parameters are correlated, analyzed, and filtered out in order to deduce contextual situations. Then, based on the knowledge that has been defined by the network operator (in the form of rules and information models) or built through cognitive tasks, a CNM makes a decision about the most appropriate reaction. In our case, we choose to tackle these two stages with Fuzzy Logic, a multi-valued reasoning technique; this bootstraps both decision making and knowledge integration (induction and building procedures). Finally, the chosen reaction is enforced, calculating the reconfiguration parameters in dedicated modules. Then the reconfiguration command (channel reallocation, handover) is pushed to the network element controller.

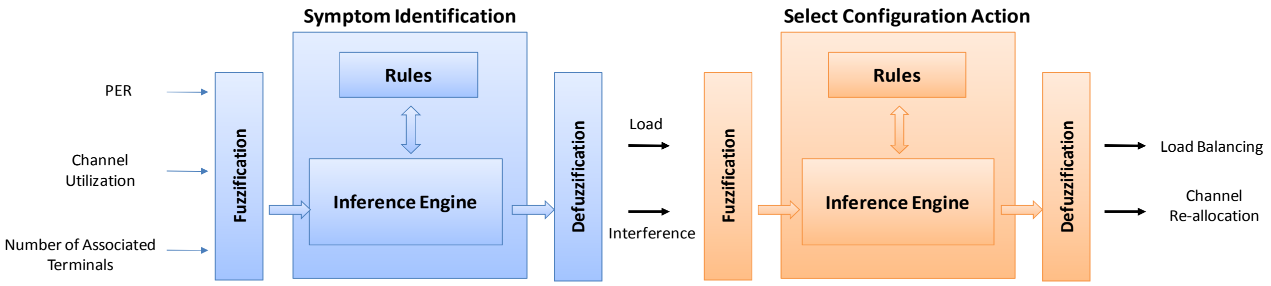


Figure 3-10: Fuzzy logic-based inference system

3.5.3 Mechanisms for Cognition Empowerment

FL enables CNMs to analyze situations and to decide appropriate reconfiguration actions. We have instantiated the three parts of a FL controller, namely the fuzzifier, the FL inference system and the defuzzifier, to fit our test case (Figure 3-10).

3.5.3.1 Knowledge Building and Decision Making

The fuzzifier undertakes to transform (fuzzify) the input values (crisp values) to the degree that these values belong to a specific state (e.g., low, high); such mapping is described by membership functions for each input variable. The membership function of each input variable captures the degree of each conceivable state for this input. To evaluate the radio interference and the traffic load, we consider as inputs three parameters monitored by the NECMs and we define their corresponding membership function: the channel utilization (low, medium, high), the packet error rate (low, medium, high), and the number of associated terminals (very low, low, high, very high). The measured input is being mapped to a specific degree for all the membership functions of this input. The same approach is used for the output variables, namely the interference and the load values. Figure 3-11 presents the range of the defined states for some of the input and output variables.

Then, the FL inference system is in charge of correlating the inputs and the outputs using simple "IF…THEN…" rules. Each rule results to a certain membership function degree for every output. The rule building procedure is based on expert's knowledge. Thereinafter, the output degrees for all the rules of the FL inference phase are aggregated. Afterwards, the FL decision engine defuzzifies the aggregated output degrees and concludes to the status of the decision maker. In our case, the result of this process corresponds to: the network element is x% loaded and the radio link is y% interfered (Figure 3-10).

NECMs and NDCMs embed the membership functions and the FL rules that we have defined to assess radio interference and high load symptoms. Operators' knowledge is integrated in all three parts of a FL inference system, and it is related to the membership functions' shape as well as to the inference engine's rules. After the calculation of the current network element's interference and load level, the second phase of the decision making module undertakes to select the proper configuration action. The second FL controller maps the output status levels of the first controller (interference and load level) to the available configuration actions (channel re-selection and assisted handover), which are being captured by the respective behaviors (Figure 3-10). The increase of the number of inputs and the corresponding increase of the required number of rules might raise significant scalability and complexity issues. The usage of a Hierarchical Fuzzy Controller, with two or three levels, could address this problem leading to the reduction of the number of the used rules [109].

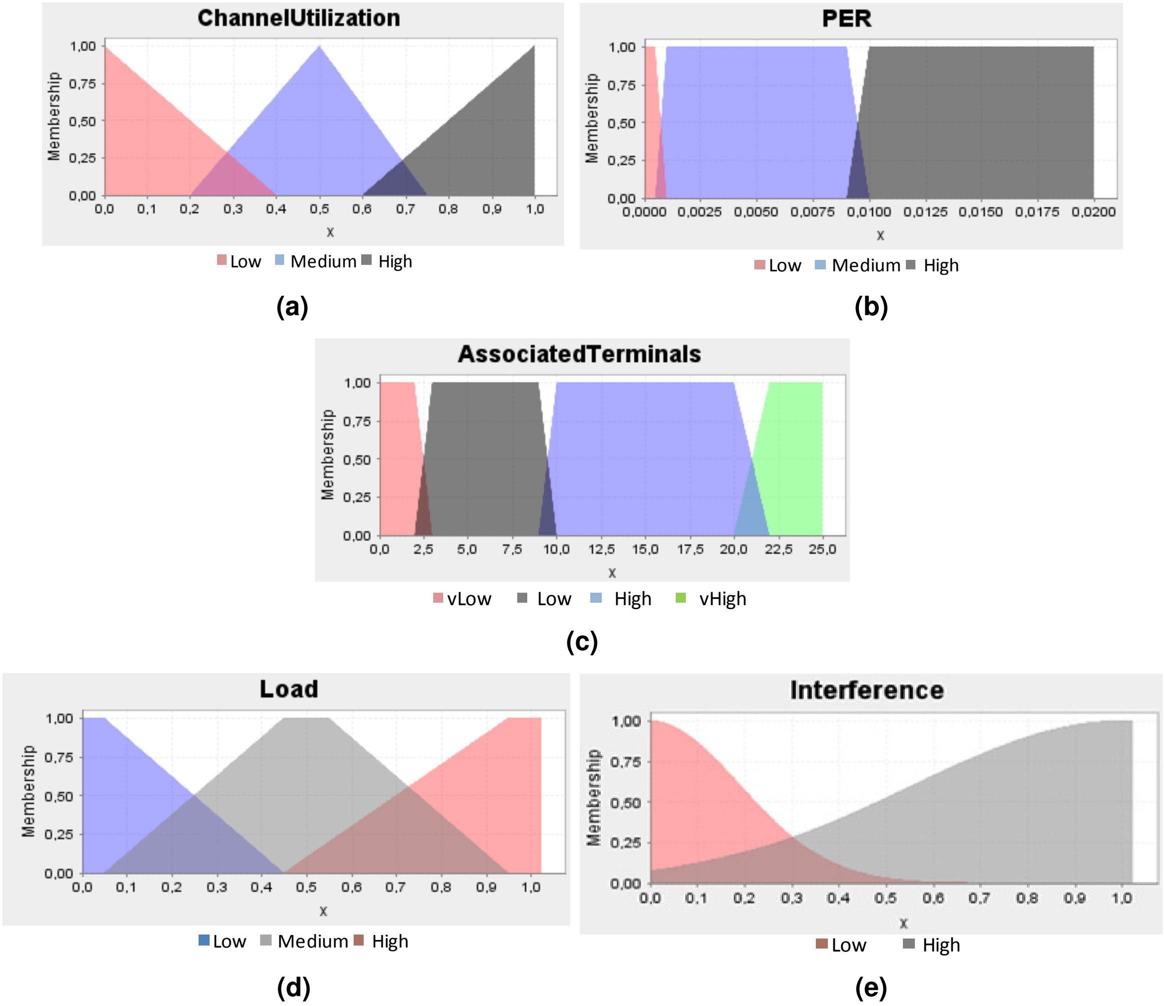


Figure 3-11: (a) Input membership functions of channel utilization, (b) Input membership functions of Packet Error Rate, (c) Input membership functions of Number of Associated Terminals, (d) Output membership functions for Load level estimation, (e) Output membership functions for Interference level estimation

3.5.3.2 Behavior Modules

The calculation of the reconfiguration parameters is performed in dedicated behaviors. The first behavior concerns channel selection. The aim is to allocate to the problematic AP the optimum channel so as to have as low interference as possible. The interference is related to the channel overlap among the neighboring WiFi channels on the one hand and to the usage degree of these channels on the other. The NECM of the AP, which has detected high interference status select a new channel using the following scheme:

- Identifies neighboring APs and the list of available channels.
- For each candidate channel (Ch_j) the NECM calculates the estimated interference cost (*IC*) by using the following objective function:

$$IC_{Ch_j} = \sum_{i=1}^{n} w_i Overlap(Ch_i, Ch_j) \tag{3-1}$$

where:

Ch_i: the channel of AP *i* that is sensed by the AP *j*,

n: number of neighboring APs,

w_i: the number of users that are associated at the neighboring AP *i*,

$Overlap(Ch_i, Ch_j)$: the level of channels *i* and *j* interference overlap.

- The NECM selects channel Ch_j that creates the minimum IC, which consequently means that the less possible interference/overlapping will be caused in the local area. Once the new channel has been selected, the configuration is enforced through the NEC.

The second implemented behavior is executed by the NDCM, when a high load symptom has been identified by an AP or BS NECM. The NDCM collects the necessary information so as to decide on the new allocation of the terminals that are attached to the heavy loaded NECM. The decision for the allocation of users to radio access technologies is based on a modified approach of the fittingness factor as it is presented in [110]. The NDCM calculates the fittingness factor for all the users of the loaded AP and finds which AP is more suitable for each user. Then the domain manager forwards the decision to those users that should be handed over in order to drop the load levels below a predefined threshold. The fittingness factor is defined as follows:

$$\Psi_{k,l} = Q_{k,l} \cdot \delta(\eta_{NF}) \tag{3-2}$$

The two parts of the equation (2) represent the user centric suitability ($Q_{k,l}$) and the network centric suitability ($\delta(\eta_{NF})$). The user centric suitability reflects the degree that every AP/BS is suitable for each user's needs. In this experimentation study we consider that such metric is the Received Signal Strength (RSS) ($Q_{k,l} = RSS_{k,l}$, $RSS_{k,l}$ is the signal strength received by the *k*-th terminal from the *l*-th AP).

The network centric suitability reflects the degree that each WiFi AP or WiMAX BS is suitable for each terminal from a network's perspective. Such metric is related to the overall and the non-flexible load (i.e., represents the requirements of the terminals that cannot connect to other APs or Radio Access Technologies) of each NECM. An empirical definition of δ(ηNF) is proposed in [110] and it is given by the following function:

$$\delta(\eta_{NF}) = \begin{cases} 1, & if\ \ \eta \le 1 - \min(\eta_{NF}, D)\ \text{OR traffic is non-flexible} \\ \left(\dfrac{1-\eta}{\min(\eta_{NF}, D)}\right)^2, & if\ \ \eta > 1 - \min(\eta_{NF}, D)\ \text{AND traffic is flexible} \end{cases} \tag{3-3}$$

where:

η is the normalized load,

η_{NF} is the normalized non-flexible load,

D is a parameter that captures the percentage of the overall bandwidth that is going to be reserved for the non-flexible load.

The NDCM calculates the fittingness factors for every terminal allocated to the loaded AP/BS in order to select the most suitable AP/BS for the terminals' re-allocation. Then, the NDCM forwards the decision to the terminals through the corresponding NECMs. The rationale behind the use of such empirical function is related to the fact that if an AP is heavily loaded with non-flexible load, it is less attractive for flexible load, thus leaving room for non-flexible users.

3.6 Performance Evaluation

We have conducted experiments to evaluate the embodied cognitive network architecture in the environment described in Section 3.3/3.4. As mentioned above, the inference engine captures two network symptoms: channel interference and high load situation. Such decisions are based on measurements of packet error rate, channel utilization, and number of associated terminals. For instance, Figure 3-12 a illustrates the variation of the interference perception, using the FL controllers, in relation to the

Table 3-1: NECM Fuzzy Logic Decision Rules

Rule	Conditions			Outcome	
	PER	**Channel Utilization**	**Terminals Number**	**Interference**	**Load**
1	Low	High	Very Low	Low	Low
2	Low	Low	Low	Low	Low
3	Low	Medium	Low	Low	Low
4	Low	High	Low	Low	Medium
5	Low	Low	Very High	Low	Medium
6	Low	Medium	Very High	Low	High
7	Medium	High	Very High	High	High
8	Medium	Low	High	Low	Medium
9	Medium	Medium	High	High	Medium
10	Medium	High	High	High	High
11	Medium	Low	Very Low	Low	Low
12	Medium	Medium	Very Low	Low	Low

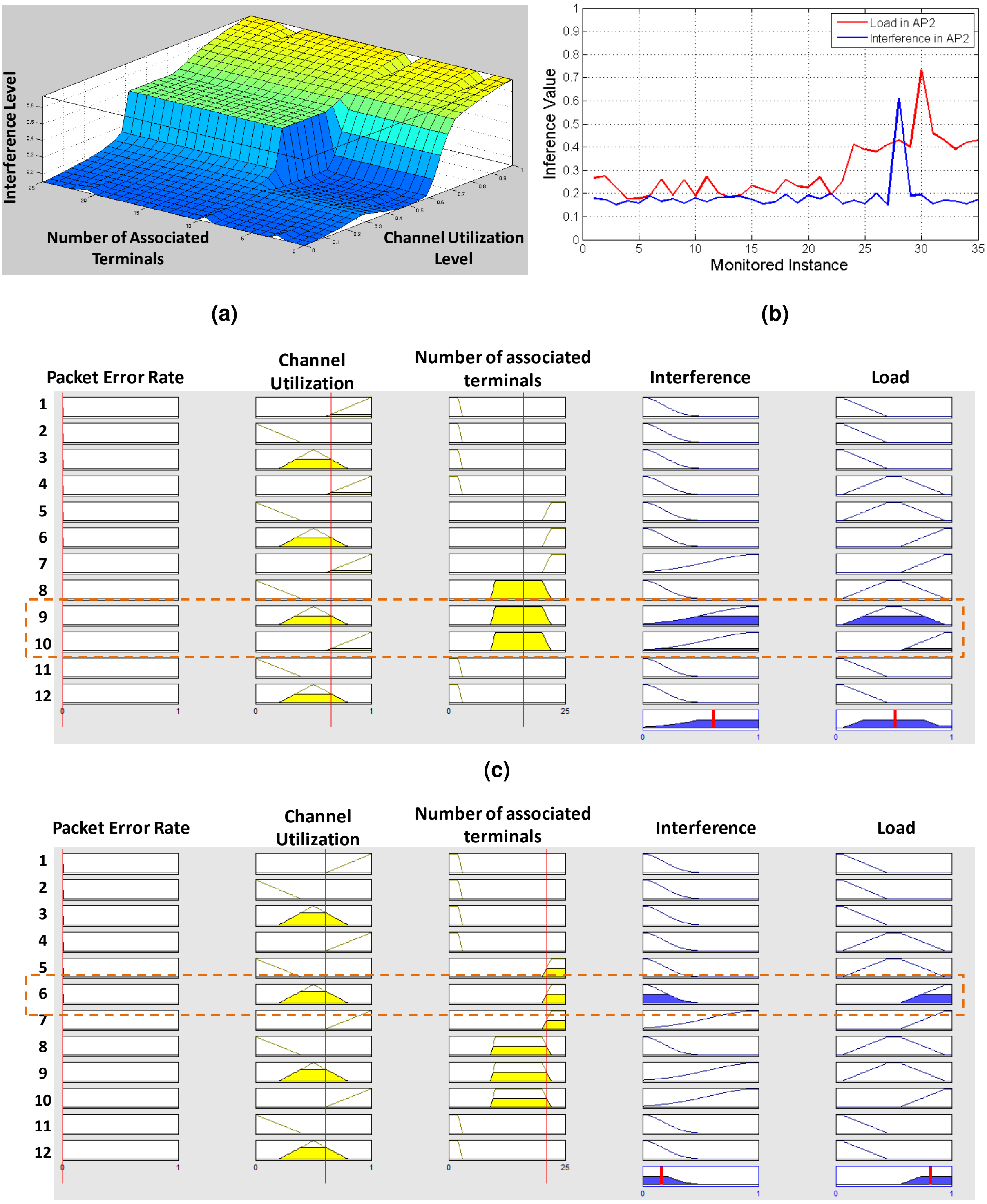
channel utilization and the number of attached terminals, having as parameter the packet error rate (i.e., set as 10^{-6} - low). The figure shows that, if an AP has medium packer error rate level (Figure 3-11-b), it identifies high interference if the channel utilization is above 0.6 and the number of associated terminals is higher than 10.

In the experimentation scenario, we have deployed the 4 APs in an office environment, which is also under the coverage area of the WiMAX BS. The initial channel allocation of the AP1, AP2, AP3 and AP4 is channel 1, 3, 5, and 7, respectively. At first, considering the users associated to each AP, the NECMs do not detect interference problems. Figure 3-12-b captures the interference level evolution for a specific time period of our experimentation for AP2. For this execution of the experiment, the NECM identifies low interference degree for all but one monitored instances of AP2. The NECM of the AP identifies high interference situation at time T27. It thus activates the channel reallocation behavior**.** Taking into account the existing traffic conditions and the overlaps with neighboring APs, this behavior estimates that the selection of channel 11 will reduce intra-channel interference for AP2. Once the decision making and re-configuration procedures have taken place, the interference level falls back to a normal level. Figure 3-12-c plots the fuzzy rules derived from expert knowledge in the experiment. The conditions and the actions that form the decision rules of Figure 3-12-c are described in Table 3-1. The membership degrees are presented for the parameter values at instance T27 (Figure 3-12-b). The two rules that are activated at T27, when the NECM identifies a high interference situation, are highlighted using a dotted circular line. The right side of the figure shows how the interference is assessed by the NECM. The monitored parameter values (packet error rate: 0.00107, channel utilization: 0.65, number of associated terminals: 16) result in a high interference degree (0.607). The total time for the NECM of AP2 to address this high interference event is 5.89 seconds, which is decomposed to a) the monitoring phase: 1.27 seconds, b) the decision making phase for problem identification and configuration action selection: 0.39 seconds, and c) the execution phase for the calculation of the appropriate channel as well as the consequent re-allocation procedure at AP2 and the associated terminals: 4.23 seconds.

Afterwards, due to the arrival of more users in the office environment and the increase of required bandwidth, the NECM of AP2 identifies a high load situation (0.723) (measured Packet Error Rate: 0.00088, Channel Utilization: 0.65, number of associated terminals: 21) at T30 (Figure 3-12-c). Consequently, the NECM of AP2 triggers the assisted handover behavior at the NDCM, which decides the re-allocation of user terminals among the available APs and the WiMAX BS (Figure 3-12-d). More specifically, the updated allocation of terminals is as follows: 9 terminals to AP1, 12 to AP2, 8 to AP3, 4 to AP4, 11 to the WiMAX BS. The allocation of associated terminals before the handover enforcement was: 8 to AP1, 21 to AP2, 7 to AP3, 3 to AP4 and 5 to the WiMAX BS. Once the decision making and the handover procedures have taken place, the load level falls back to normal. In this case, the cognitive cycles of both AP2 and the NDCM addressed cooperatively the high load event in 8.04 seconds. This duration consists of a) the monitoring phase: 1.92 seconds, b) the decision making phase for problem identification and configuration action selection: 0.72 seconds, and c) the execution phase for the calculation of the terminals to be handed over as well as the handover procedure from AP2 to other available APs and the WiMAX BS: 5.40 seconds.

3.7 Evaluation of Deployment

One of the main reasons for the embodiment of cognition in network elements is to address the increasing complexity of broadband wireless networks management. The system architecture that is described in this section and the corresponding deployment and experimentation activities prove that cognitive network management is a tangible paradigm that enables the smart reallocation of network resources and the management



(a) (b) (c) (d)

Figure 3-12: (a) Interference surface, (b) AP2 NECM interference and load levels calculation, (c) AP2 NECM fuzzy logic rules activation for high interference case, (d) AP2 NECM fuzzy logic rules activation for high load case

of the user traffic according to the network conditions. Apart from the advantages that the CNMs provide both to the network operator and indirectly to the end user, there is a multilateral cost that is introduced in order to perceive, analyze and address network complexity, which is considered inevitable.

The operation of the cognitive manager per network element increases the average utilization of processing resources as well as the average memory usage of the network device, where the NECM is incorporated (AP, BS). For the deployment and the test case described in the previous section, the average memory usage of the network node appears an increase of 15-20%, while the average Central Processing Unit (CPU) appears an increase of 13-18%, when the NECM is periodically enabled. The CPU and memory usage measurements include both the periodic operation of each Cognitive agent as well as the problem solving tasks of an NECM when a problem is identified (e.g., channel re-allocation due to high interference, vertical assisted handover due to high load status). The overhead in the memory and CPU utilization leads to the increase of the energy that is consumed by each device, increasing the total cost for the operation of the network. In some cases (e.g., for home or office devices), additional hardware resources will be required in order to support the NECM or NDCM tasks.

Another important aspect is the time taken by CNMs to perform a whole cognitive cycle (i.e., monitor, decide, execute). Based on the conducted experiments we observe that the bottleneck points, in terms of delay for an NECM to detect a problematic situation and thereinafter solve it, are the monitoring and execution phases. They consume 23% and 69% of the time to solve a problem, correspondingly. The involvement of the network communication phase for the interactions among cognitive agents and the actual time to perceive the situation by using the FL inference system is in average 8% of the total delay. The duration of the execution phase is affected by several technical and implementation-related issues, especially when proprietary/legacy tools have been used, since further translation and mediation layer is required. There is the necessity for the specification of cross layer monitoring and execution interfaces that will reduce cognitive cycle time and implementation effort.

Knowledge exchange among cognitive managers introduces additional traffic into the network. The cost emerging from the cognition deployment is thus affected by the graph topology and the number of neighboring NECMs and associated terminals. We observed that the denser the topology is, the higher is the duration for fault identification and for the problem solving phase. Consequently, the density of the network graph affects the CPU and memory cost. One possible solution consists in adapting the periodicity of cognitive cycle checks according to the dynamicity of the network environment. Another key requirement for the efficient performance of the cognitive managers is the fine tuning of its cognitive capabilities, that is, in our approach, the appropriate selection of the thresholds for faults identification and the parameters or weights of the objective functions that each behavior uses. The administrator of the cognitive managers usually sets the initial values of these parameters or thresholds based on accumulated experience. However, the situation and the different characteristics of the network segments may lead to different values for every device. Even if the initial selection of parameters per network node is correct, continuous updates during their digital life might be required. Thus, the cognitive manager should have the capability for on-line and/or off-line tuning of network parameters (e.g., High load threshold, number terminals weight for Channel selection objective function) by exploiting the feedback from previous events or from historic data that are stored locally.

Finally, the design principles of modularity, genericity and extensibility constitute a solid basis for the development of distributed cognitive managers. However, a global configuration interface for a CNM should be considered in order to easily configure and deploy it. A typical example is the introduction of a new operator's rule that has to be integrated into all the running CNMs. In addition, such a configuration Application Programming Interfaces (API) would provide a remote start and stop facility to go towards the optimization of the fingerprint of the different modules.

Autonomic and cognitive network management systems performance analysis is in its infancy. There is no well-defined and commonly agreed measurement methodology to assess network self-management performance. The comparison of self-managed systems is not an easy task, since different (or partial) functionalities of self-management architectures have been implemented in each experimentation attempt. Thus, the published results of existing studies cannot be generalized. Each experimentation approach implements a different use case/scenario, while the various configuration of the network devices or the network topology is not provided in many cases. All the above make difficult the direct comparison. However, experimental research remains a good starting point in generating knowledge and validating proposed self-management approaches.

CHAPTER 4

CLUSTER-BASED STRUCTURE OF SELF-ORGANIZING NETWORKS

In SONs, the communication system is considered as an autonomic element, which is capable of monitoring its network-related state and modifying it based on policy rules or business goals that administrators have specified. The cognitive control loop, known also as feedback control loop, is a necessary enabler for in-network management realization that leads to its autonomy. As mentioned in the previous chapter, the control loop includes processes for monitoring and perceiving the network node's internal state and environmental conditions, and then planning, deciding, and finally adapting according to the identified conditions. The points where control loops are placed, as well as their way of interacting, is a key design issue for the operation of network management and the performance of management tasks.

The adoption of a totally centralized solution, on one hand, has the advantage that the management entity will have global view of the network status, which facilitates optimal decision making, but on the other hand there is a single point of failure, where scalability, computational, local optimization and local search issues arise. On the other hand, a fully distributed solution requires continuous coordination of control loop decisions, both spatially and temporally, in order to avoid conflicts or ripple effect phenomena. For that purpose, we have adopted a hybrid approach, where the network nodes e.g., APs form clusters in order to perform their management functions e.g., coverage and capacity self-optimization, energy management.

Clusters are organization structures used for the collaborative tackling of network management problems, and they are formed following a common known scheme. Clusters facilitate the cooperation and the coordination of a group of network nodes for identifying and solving network management problems. In the literature, there are several clustering algorithms, which are mainly targeting wireless sensor networks or mobile ad-hoc networks, but the majority of them are application-specific (e.g., energy-efficient, mobility-aware). In this chapter we propose an algorithm, entitled SYSTAS, for the discovery and the establishment of clusters among network nodes, based on the features of the physical network topology [111], [112]. The density of the network graph and the preferential attachment model are used in order to form logical topologies i.e., clusters. These structures enhance the performance improvement capabilities of autonomic network management and contribute towards the more rational usage of network resources. The proposed scheme is evaluated using various graphs of access points.

4.1 Background

Clustering is multidisciplinary issue, where various algorithms have been proposed the last decades from the research community. Clustering of graphs is the task of grouping the vertices of the graph into clusters, taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters [113]. Clustering in mobile ad hoc network (MANETS) can be defined as the virtual partitioning of the dynamic nodes into various groups. Groups of the nodes are made with respect to their nearness to other nodes. Two nodes are said to be neighbor of each other when both of them lie within their transmission range and set up a bidirectional link between them [114]. In the context of MANETs or WSNs the formation of a clustering structure guarantees some basic performance achievements. For instance, the spatial reuse of network resources to increase the system capacity, routing, energy efficiency, and load balancing are some of the performance objectives of the proposed clustering schemes. In some of the clustering literature, a cluster in a graph is called a community [115]. In [114] the following categories have been identified for the classification of clustering schemes, which also specify the objective functions for the formation of the corresponding clusters:

- Dominating-Set-based (DS-based) clustering,
- Low-maintenance clustering,
- Mobility-aware clustering,
- Energy-efficient clustering,
- Load-balancing clustering,
- Combined-metrics-based clustering.

The number of hops of each cluster (i.e., single-hop, multi-hop), the existence of cluster-head, nodes' mobility, the overlap of clusters as well as the usage of control messages (active clustering, passive clustering) are some of the criteria that are used for the categorization of the clustering schemes. Apart from the performance improvement that is achieved through each clustering scheme, the efficiency of an algorithm could be also evaluated as follows:

- Message exchange (Signaling cost) for maintenance,
- Ripple effect of re-clustering,
- Number of Iterations (i.e., computation rounds),

- Communication Complexity.

To the best of our knowledge, the concept of partitioning the dynamic network into logical clusters was initially proposed by Baker and Ephremides [116]. They designed a self-starting, distributed algorithm to establish and maintain a connected architecture even with the node mobility or node failure. Thus the authors of [116] proposed a new architecture called the linked cluster architecture (LCA), where the network is organized into a set of node clusters and each node belongs to at least one cluster. Thereinafter, a variation to LCA was proposed by Ephremides et al. in [117] entitled LID algorithm. In this scheme, every node is assigned with a unique non-negative identification number (ID), which is the deciding factor for the status of a node. Each node broadcasts its ID to its neighbors and receives the same from the latter. If a node listens to all the neighbors' IDs that are higher than its own ID, then it declares itself as the cluster head among its immediate neighbors. This process is repeated till all the nodes are assigned with the role of a head or a member of a cluster. Thereinafter, a mobility metric based version of LID algorithm was proposed by Basu et al. in [118]. The algorithm, named MOBIC uses a mobility-based metric for cluster formation and the calculation of weights of the nodes in the network. The ratio of two consecutive signal strengths received by a node to know its relative motion with respect to its neighbors is used by the author for the mobility modeling. In addition, based on LCA, authors in [119] have proposed the Highest Connectivity (HC) algorithm, where cluster heads are elected those nodes that have the highest degree of connectivity. In [120] Das et al. proposed a weighted clustering algorithm (WCA) where a set of node parameters such as degree of connectivity, mobility, transmission power and available battery power are taking into account for the selection of a cluster head and are given different weights depending on the network scenario. Basagni et al. presented a distributed clustering algorithm (DCA) [121] and the mobility adaptive clustering algorithm (DMAC) [122]. In DCA each node is associated with a unique parameter (i.e., weight) that decides the role of a node. The weights may be the function of node transmission range or node mobility. In [123] authors propose Least Cluster Change (LCC), which enhances LID algorithm. In LCC the clustering algorithm is divided into two steps: cluster formation and cluster maintenance. The cluster formation is based on LID, (i.e., initially nodes with the lowest ID in their neighborhoods are chosen as cluster-heads). Re-clustering (i.e., maintenance) is event-driven and invoked in only two cases: a) when two cluster heads move into the reach range of each other, one gives up the cluster-head role, b) when a mobile node cannot access any cluster-head, it rebuilds the cluster structure for the network according to

LIC. The authors in [124] have proposed Hierarchical control clustering (HCC) to create a hierarchical control structure for multi-hop wireless networks. HCC cluster formation is based on a BFS (Breadth First Search) tree, which involves constructing a spanning tree in time proportional to the diameter of the network by doing a distributed breadth-first search. In addition, an algorithm for cluster establishment (ACE) [125] based on the node degree parameter has been proposed for WSNs. ACE uses a fixed number of iterations for the election of heads; during each iteration a node is allowed to assess its potential as a head before becoming real one and stepping down if it is not the best cluster-head at the moment. When a node reaches the set number of iterations, it decides based on the available information. The sensor node elects itself as a head if it detects that many nodes in its neighborhood do not belong to any cluster. It broadcasts message to invite its neighbors to join it. Moreover, cluster formation algorithms have been proposed for energy saving purposes especially for WSNs. For instance, Low energy adaptive clustering hierarchy (LEACH) [126] is based on the strength of received signal, where cluster-heads serve as routers to the base-stations and all the data fusion and aggregation are performed locally. There are also several extensions of LEACH e.g., HEED [127]. Power-Efficient and Adaptive Clustering Hierarchy (PEACH) [128] has been proposed in order to reduce the consumed energy during cluster formation in WSNs.

A more detailed analysis and comparison of the different algorithms that have been proposed for MANETs or WSNs is available in [111], [129] and [130], while, Table 4-1 summarizes the main characteristics of the above mentioned clustering schemes.

Table 4-1: Clustering schemes Comparison [129]

Clustering Schemes	Cluster Head Characteristics					Clustering Characteristics		
	Existence	Count	Selectivity	Role	Mobility	Hops	Control message	Overlapping
HCC	Yes	Variable	Pre-assign	Aggregation &Relaying	Possible	multi-hop	Yes	Yes
LEACH	Yes	Variable	Random	Aggregation &Relaying	Fixed BS	1-hop	Yes	No
PEGASIS	No	N/A	N/A	One for Aggr. &Relaying	Possible	1-hop	Yes	Possible
HEED	Yes	Variable	Random	Aggregation & Relaying	Stationary	multi-hop	Yes	No
TEEN	Yes	Variable	Random	Aggregation & Relaying	Possible	1-hop	Yes	Possible
APTEEN	Yes	Variable	Random	Aggregation & Relaying	Possible	1-hop	Yes	Possible
ACE	Yes	Variable	Random	Aggregation & Relaying	Possible	multi-hop	Yes	Very Low
EECS	Yes	Constant	Random	Aggregation & Relaying	Possible	1-hop	Yes	N/A
EEUC	Yes	Variable	Proportional	Aggregation & Relaying	Possible	multi--hop	Yes	N/A
PEACH	Yes	Variable	Probabilistic	Aggregation & Relaying	N/A	multi-hop	Yes	Possible
FLOC	Yes	Variable	Random	Aggregation & Relaying	Possible	multi-hop	Yes	Minimum
LID	Yes	Variable	Lowest-ID	Aggregation & Relaying	Possible	multi-hop	Yes	No
DCA	Yes	Variable	Bigger Weight	Aggregation & Relaying	Possible	1-hop	Yes	No
3hBAC	Yes	Variable	Lowest-ID	Aggregation & Relaying	Possible	1-hop	Yes	No
CDS	Yes	Variable	Minimum number	Dominating	Possible	1-hop	Yes	No

Clusters discovery has been also studied in the context of graph theory. The previous work on methods for discovering groups in networks is divided into two principal lines of research [131].

- Graph partitioning (GP), has been pursued particularly in computer science and related fields, with applications in parallel computing and integrated circuit design, among other areas.
- The second, identified by names such as block modeling, hierarchical clustering, or community structure detection (CS), has been pursued by sociologists and more recently by physicists, biologists, and applied mathematicians, with applications especially to social and biological networks.

In GP we know, or we have an indication of the number of groups, while in CS detection methods, we assume that the network of interest divides naturally into subgroups and the experimenter's job is to find those groups. The number and size of the groups are thus determined by the network itself and not by the experimenter.

There are various structural metrics that could be used in order to describe the properties of a graph topology e.g., degree (number of edges incident to the node), degree distribution, path length, clustering coefficient, centralities, geodesic distance.

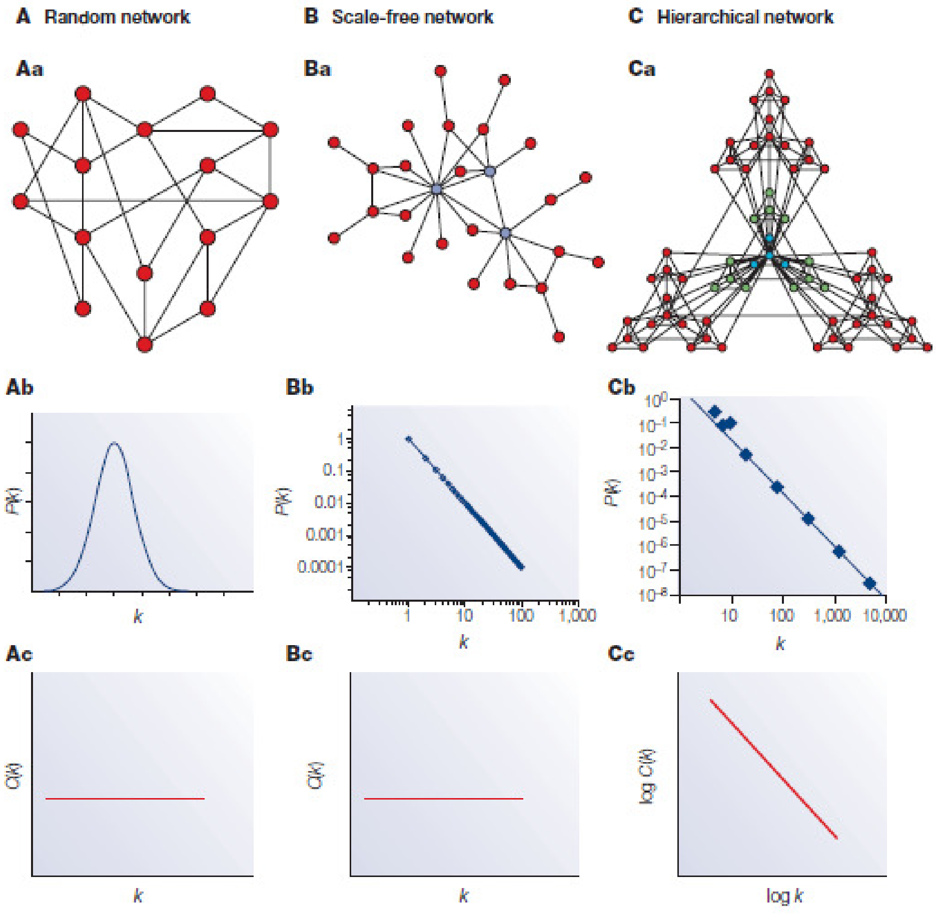


Figure 4-1: Structural models and structural metrics properties [132]

According to these properties four main categories of structural models could be identified:

- Random network: a fixed number of nodes are connected randomly to each other.
- Small-world network: Any two nodes can be connected with a path of a few links only.
- Scale-free network: The network's properties are determined by hubs.
- Hierarchical network: seamlessly integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution with degree exponent $\gamma = 1 + \ln 4/\ln 3 = 2.26$.

Each structural model has specific characteristics as regards the values of the key structural metrics e.g., degree distribution (P(k)), clustering coefficient (C(k)). Figure 4-1, based on Barabasi and Oltvai work [132] presents the properties of three main network models. In a random network, the node degree follows Poisson distribution (Figure 4-1, Ab), while the clustering coefficient is independent of node's degree (Figure 4-1, Ac). Scale-free networks are characterized by a power-law degree distribution (Figure 4-1, Bb).

4.2 SYSTAS: Density-based Algorithm for Clusters Discovery in Self-Organizing Networks

The application of the proposed algorithm leads to the election of the head and the specification of the borders of the clusters through the allocation of the member nodes to the elected heads. All nodes have the capability to discover their physical topology and then, using the common known scheme, to form, in a distributed way, the appropriate number of clusters. At this point we should note that the number of elected heads defines the number of the formed clusters. Clusters are non-overlapping and consist of two types of nodes: a) the simple member node, and b) the head node (Figure 4-2). The proposed clustering algorithm considers the low or zero mobility levels of the nodes and it consists of the bootstrapping phase and the maintenance phase. The proposed algorithm is based on the topological characteristics of the network area (nodes degree, clustering co-efficiency). Heads are elected and consequently clusters are formed based on the network density and by a process of "preferential attachment", where nodes prefer to join to the more popular existing members. Two nodes are considered neighboring if they are within each other's transmission range.

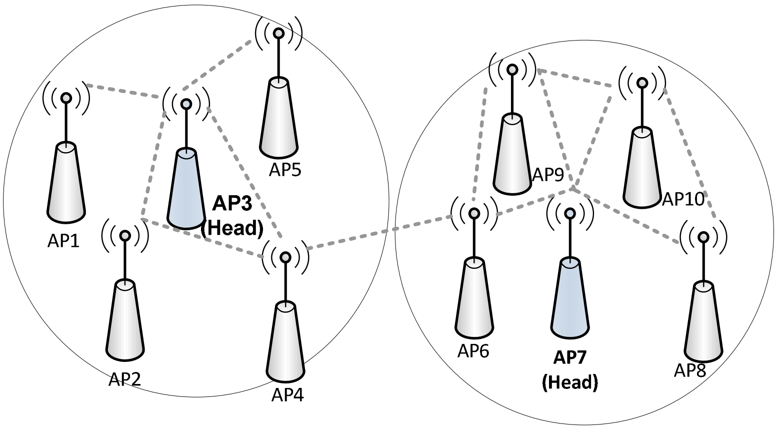


Figure 4-2: Sample topology – Clusters of APs

The number of deployed nodes (e.g., APs) as well as the overlapping of their transmission ranges are taken into account. We assume that a network topology is represented by a connected and undirected graph $G = (V, E)$, where V is the set of network nodes in the topology and E is the set of edges (connectivity links) between network nodes. An edge exists between two nodes (e.g., APs), if one is within the coverage area of the other. Let the cardinalities of V and E be denoted by n, and e, respectively; i.e., n is the number of nodes that constitute the network topology, while e is the number of the existing connectivity links among the nodes of the network topology. The input variables of the proposed algorithm are:

a) *adv*: the maximum number of hops that each head node advertises its presence.

b) *mF*: the maximum number of member nodes that triggers a cluster to be merged with a neighboring cluster.

c) R_{thr}: the threshold of inter-cluster and intra-cluster edges ratio (denoted by R) that initiates a cluster merging process.

4.2.1 Bootstrapping Phase

1. Firstly, each node i (i.e., AP) monitors its proximity and specifically one-hop away neighbours and calculates its degree D_i. The D_i metric denotes the number of one hop away neighbours of node i.

2. Each node i advertise to its one-hop away nodes the calculated D_i.

3. After receiving all neighboring nodes' degrees, each node elects its head (H_j). The node with the largest degree, including the degree of itself, is elected as the head of node i. The node that is head, even for a single node, is not considered as a member node of other clusters. Each member node stores the degree and the ID of the elected

head, while a head keeps the number and the ID of the member nodes that constitute its cluster. In this step the first set of clusters are formed according to the degree of each node.

4. All head nodes that consist of a zero number of member nodes are merged with the neighboring cluster that has the largest degree.

5. Each head node H_j calculates the existing number of its member nodes denoted by M_j, ($M_j \leq D_j$).

6. Each head node advertises TTL hops away its M_j value (The initial value of the time-to-live field is one (TTL=1)).

7. Each member node that receives an advertisement message, undertakes to forward the message to a neighboring member node if the following conditions are met:

- $TTL-1 \neq 0$
- It belongs to the same cluster with the previous node (member node or head node) that has sent the respective message.

A head node does not forward the advertisement messages of other heads. The goal of the above conditions is to avoid the creation of disconnected cluster areas.

8. Each member node collects all advertisement messages that has received from the elected heads that are in maximum TTL hops away and calculates the influence factor of each head. The influence of each head H_k on a member node i, denoted by IF_i^k is calculated as follows:

$$IF_i^k = \begin{cases} M_k, & if \quad TTL = 1 \\ S_{i,k}, & if \quad TTL > 1 \end{cases} \tag{4-1}$$

where M_k is the number of members of the head AP k and $S_{i,k}$ denotes the total number of advertisement messages that member node i has received from the head node H_k.

Each member node selects the head (i.e., cluster to join) with the largest IF value. Hence, member APs following the preferential attachment model, join those clusters (i.e., head nodes) that appear higher density and potentially affect their operation or communication tasks.

9. The clusters that consist of a very small number of members (e.g., $mF=1$) are merged with the neighboring cluster that has the largest dominance factor, denoted DF. The DF

metric of each neighboring cluster (*g*), takes into account the number of inter-cluster edges ($I_{j,g}^{e}$) between the under merging cluster *j (i.e.,* the head and member nodes) and the neighboring one (*g*). Hence, *DF* is calculated as follows:

$$DF_j^g = \frac{I_{j,g}^{e}}{I_j^e} \tag{4-2}$$

where I_j^e denotes the total number of inter-cluster edges of cluster *j*. Consequently, each cluster selects to merge with the neighboring cluster that has the largest *DF* value. This merging phase facilitates the uniform extension of the advertisement space of each head without being blocked from too small (local) clusters.

10. Each head node re-calculates the number of its member nodes, updating M_j. If *TTL< advH* then *TTL = TTL+1* and GOTO step 6.

11. After the completion of the advertisement phase, each head calculates the inter-cluster and intra-cluster edges ratio of its cluster (R_j). Metric R_j is calculated as follows:

$$R_j = \frac{I_j^e}{I_j^a} \tag{4-3}$$

where I_j^a denotes the total number of intra-cluster edges of cluster *j*. If $R_j > R_{thr}$ then cluster j, regardless of its constituent number of members, triggers its merging with the neighboring cluster that has the maximum *DF* value, using (4-2). The rationale of this merging phase through metric *R* is to avoid having clusters which nodes are not clearly independent in term of interactions (i.e., edges) from neighboring clusters.

The merging phase using either metric *R* or the number of member nodes leads to the reduction of the formed clusters (i.e., elected heads) and consequently to the increase of the allocated member nodes per cluster. After the end of the merging phase, the clusters have been formed and the heads of each cluster have been elected. Finally, each head node is aware of the member nodes that constitute its cluster, while each member node is aware about the ID of the head (e.g., MAC address) that is assigned to and their distance in terms of hops. Figure 4-3 illustrates the main steps of SYSTAS.

4.2.2 Maintenance Phase

After the establishment of the clusters, the maintenance phase undertakes to deals with the changes in the topology. The continuous re-organization of the formed clusters and heads re-election is not considered as an efficient action, since the communication

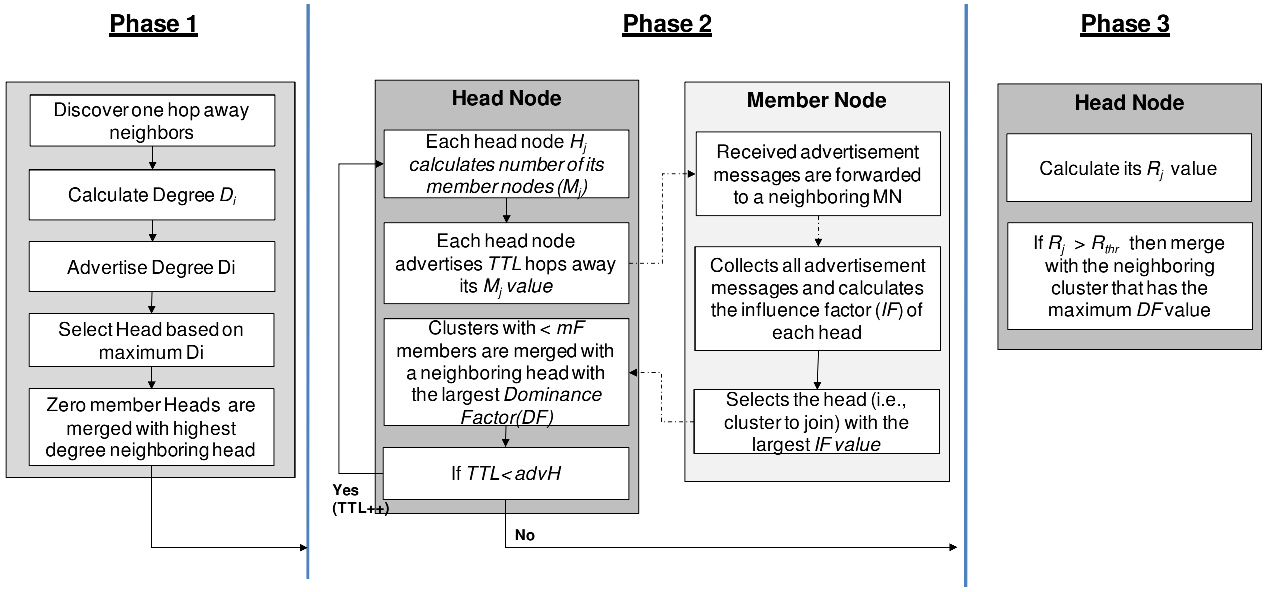


Figure 4-3: SYSTAS clustering algorithm

and computational cost for continuous updates increases the complexity; hence eliminating the advantages of established clusters. For that reason, we directly address the effect of a node switch-on (i.e., join phase) or switch-off (i.e., leave phase), while clusters fine-tuning takes place on a greater time-scale. The bootstrapping process is repeated in the case that the number nodes or the density of the network graph has significantly changed.

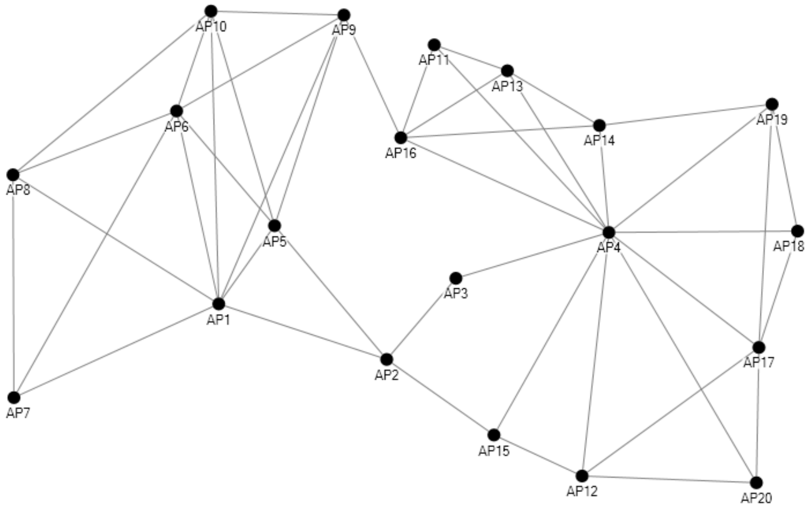
In the case of an activation of a new node (i.e., join phase), the latter informs neighboring nodes about its presence and retrieves the head ID of the neighboring (i.e., one-hop away) clusters as well as their size (M_j),. The candidate node joins the cluster with the maximum *DF* value, using (2). On the other hand, when a network node (i.e., AP) is de-activated, the actions of the neighboring nodes are affected by the specific type of the de-activated node (i.e., simple member or head). In the case of a head node, then all the simple members of the respective cluster elect as a new head the node with the largest degree in the cluster. If a simple member has been de-activated, then the cluster is not re-organized. A neighboring node that does not have any alternative route towards the head of the cluster that it belongs, undertakes to identify the next most dominant cluster to join, by using equation (4-2).

4.3 Performance Evaluation

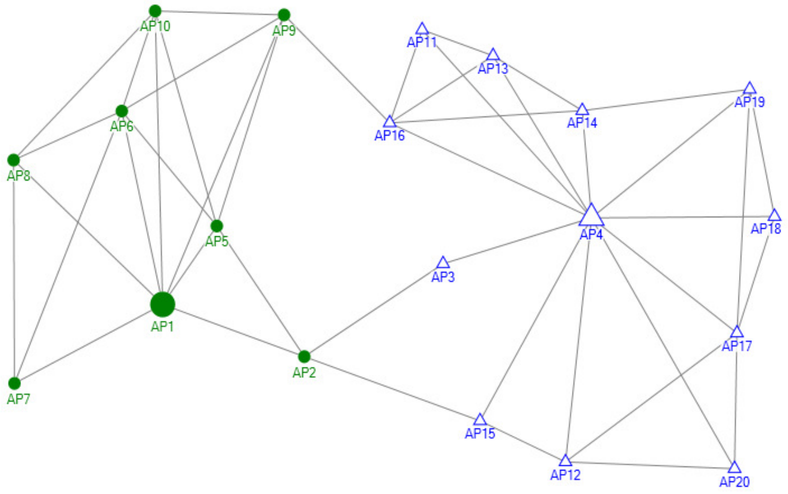
We study the proposed cluster formation algorithm using various network topologies. The graph of WLAN access points in an urban environment follow the structure of a random network topology, where nodes are connected randomly to each other (Figure 4-1). An edge between two APs denotes that they are within each other transmission range. We have set mF= 1 for adv=1, adv=2 and adv=3. The quality of the formed clusters is assessed using the modularity score Q metric that is introduced in [133]:

$$Q = Tr(e) - \left\| e^2 \right\| \tag{4-4}$$

where *e* is a symmetric matrix whose element e_{ij} is the fraction of edges in the network that connects vertices in communities *i* and *j*, and *Tr(e)* is the trace of matrix *e* (i.e., the sum of elements from its main diagonal).



(a)



(d)

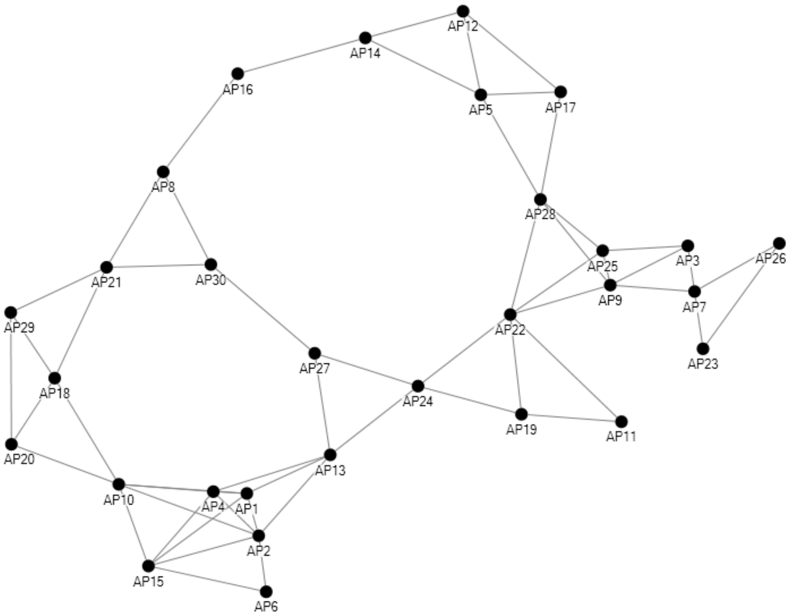
Figure 4-4: Topology of 20 nodes (a) Graph visualization, (b) Clusters visualization (adv=2, R_{thr}=1)

Figure 4-4-a illustrates a network graph that consists of 20 APs with density= 0.236 and average geodesic distance=2.29. At this point we should note that the density of a graph is calculated by using the formula that is available in [134]. By applying the cluster formation algorithm two clusters are formed and the elected heads are AP 1 and AP 4, which consist of 7 and 11 members respectively (Figure 4-4-b). Table 4-2 shows that the formed clusters for this simple topology are the same regardless of the set parameters.

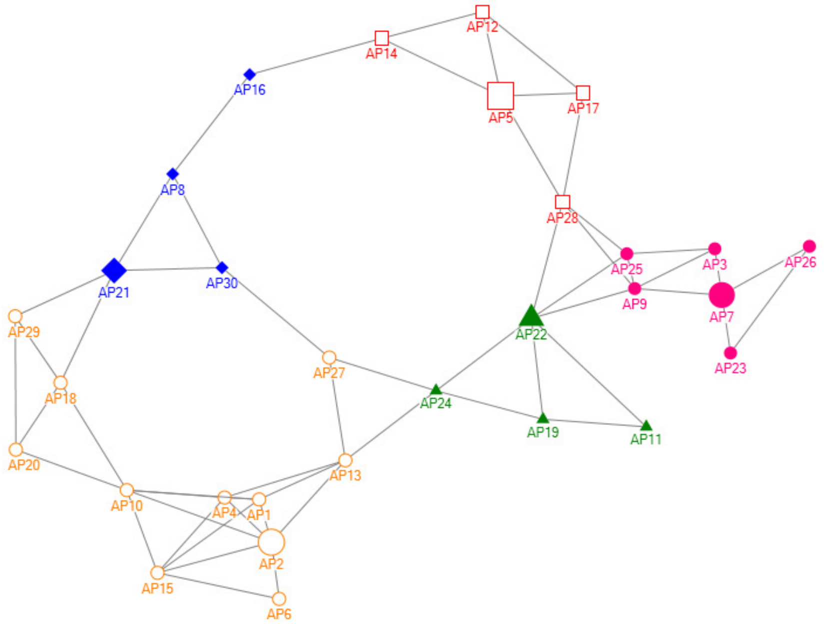
Table 4-2: Discovered Clusters for a 20 nodes topology using various input parameters

adv	R_{thr}	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	2	AP1(7), AP20(11)	0.424
1	**1**	2	AP1(7), AP20(11)	0.424
2	∞	2	AP1(7), AP20(11)	0.424
2	**1**	2	AP1(7), AP20(11)	0.424
3	∞	2	AP1(8), AP20(10)	**0.429**
3	**1**	2	AP1(8), AP20(10)	**0.429**

Moreover, five clusters have been formed applying the introduced scheme on the network graph of Figure 4-5-a, which consist of 30 APs (adv=2 and R_{thr}=1). In this case the nodes are more sparsely placed (density= 0.128, avg geodesic distance= 3.56). The elected heads are AP 2, AP 21, AP 5, AP 7, and AP 22 (Figure 4-5-d). Table 4-3 presents the clusters and the number of members that have been formed setting different values to adv and R_{thr} parameters. In this sparse graph a smaller adv value discovers more efficiently, in terms of Q, existing clusters. At this point, I would like to note that NodeXL software has been used for the visualization of the network graphs [135].



(a)



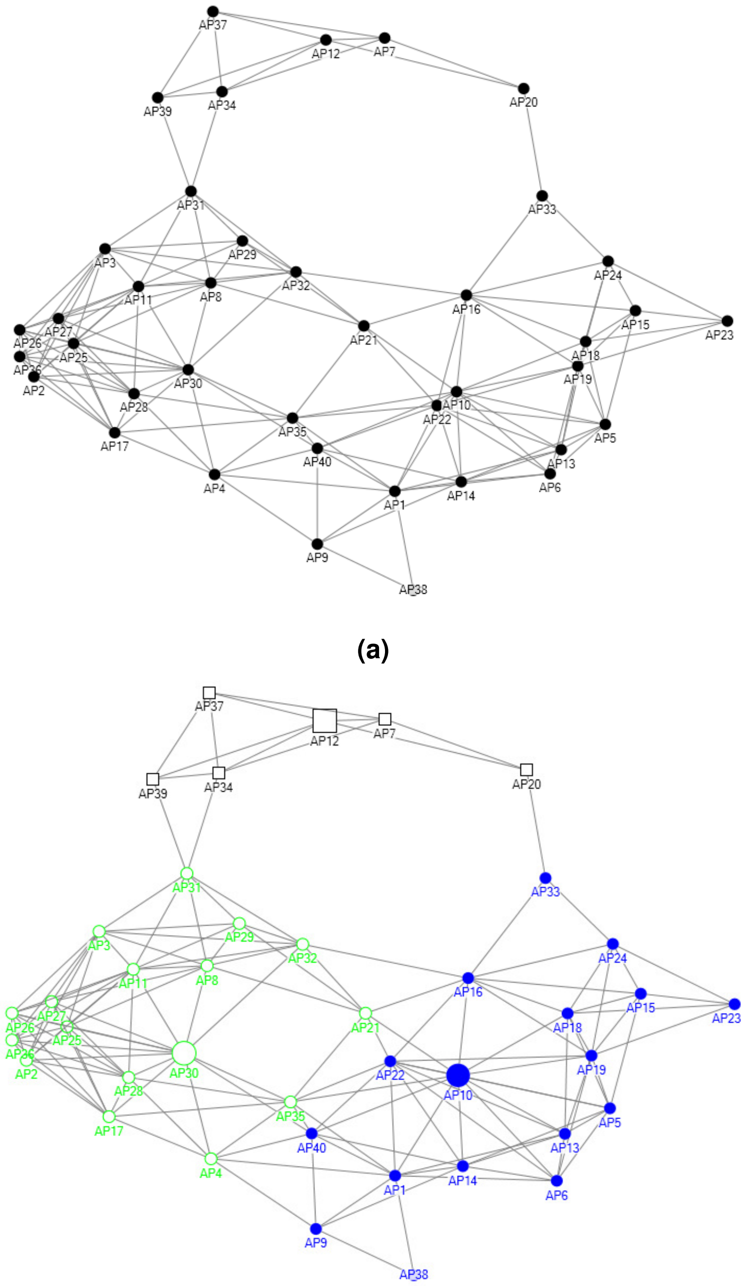
(d)

Figure 4-5: Topology of 30 nodes (a) Graph visualization, (b) Clusters visualization (adv=2, R_{thr}=1)

Table 4-3: Discovered Clusters for a 30 nodes topology using various input parameters

adv	***R_{thr}***	***No. of Clusters***	***Cluster Heads***	***Q (Modularity)***
1	∞	5	AP2(10), AP7 (4) , AP5 (4), AP21 (3), AP22 (4)	0.523
1	**1**	4	AP2(10), AP7 (9) , AP5 (5), AP21 (3)	0.528
2	∞	5	AP2(10), AP7 (4) , AP5 (4), AP21 (3), AP22 (4)	**0.540**
2	**1**	5	AP2(10), AP7 (5) , AP5 (4), AP21 (3), AP22 (3)	**0.540**
3	∞	5	AP2(10), AP7 (5) , AP5 (4), AP21 (3), AP22 (3)	0.511
3	**1**	4	AP2(11), AP7 (9) , AP5 (3), AP21 (3)	0.521

Figure 4-6-a presents the initial physical topology for a network graph of 40 APs which density is 0.198 and the average geodesic distance = 2.7. Using the proposed clusters formation scheme three clusters are formed (AP 30, AP 12, AP 10). Figure 4-6-b illustrates the constituent members as well as the intra-compartment edges of each cluster. Table 4-4 presents the clusters and the number of members that have been formed setting different values to *adv and* R_{thr} parameters.



(a)

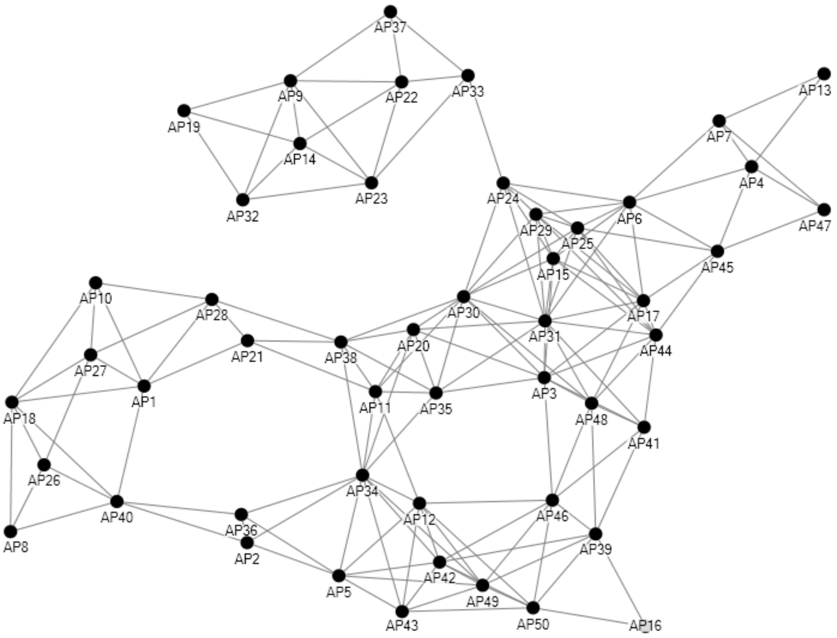
(d)

Figure 4-6: Topology of 40 nodes (a) Graph visualization, (b) Clusters visualization (adv=3, $R_{thr}=1$)

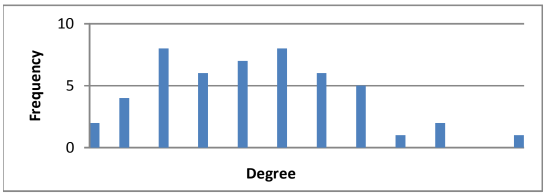
Figure 4-7-a presents the initial physical topology for a network graph of 50 APs which density is 0.128 and the average geodesic distance = 3.62. Using the proposed clusters formation scheme five clusters are discovered (AP34 (11), AP31 (14), AP6 (5), AP18 (8)). Figure 4-7-b illustrates the constituent members as well as the intra-compartment edges of each cluster. Table 4-5 presents the clusters and the number of members that have been formed setting different values to *adv and* R_{thr} parameters.

Table 4-4: Discovered Clusters for a 40 nodes topology using various input parameters

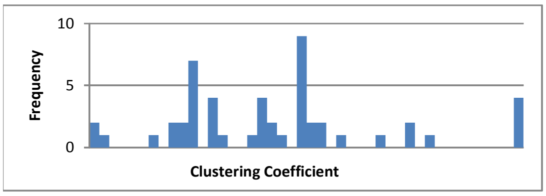
adv	*R_{thr}*	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	4	AP10(10), AP30 (16) , AP12 (5), AP19 (5)	0.372
1	**1**	3	AP10(16), AP30 (16) , AP12 (5)	0.448
2	∞	5	AP10(6), AP30 (17) , AP12 (5), AP19 (5), AP1(2),	0.404
2	**1**	3	AP10(15), AP30 (17) , AP12 (5)	0.449
3	∞	3	AP10(16), AP30 (16) , AP12 (5)	0.468
3	**1**	3	AP10(16), AP30 (16) , AP12 (5)	**0.468**



(a)

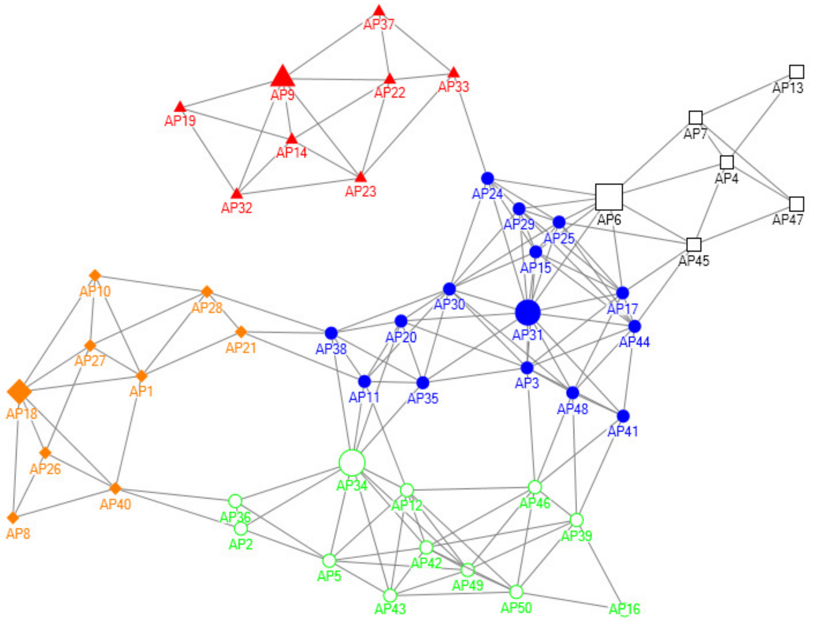


Minimum Degree	2
Maximum Degree	13
Average Degree	6.280



Minimum Clustering Coefficient	0.333
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.623

(b) **(c)**



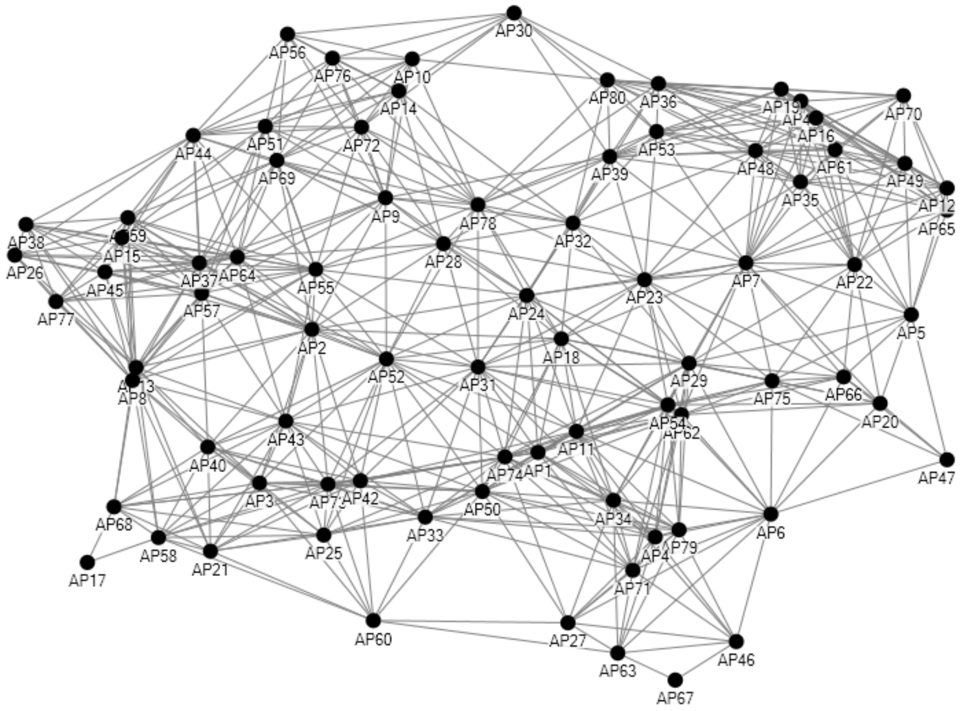
(d)

Figure 4-7: Topology of 50 Nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters visualization (adv=2, $R_{thr}=1$)

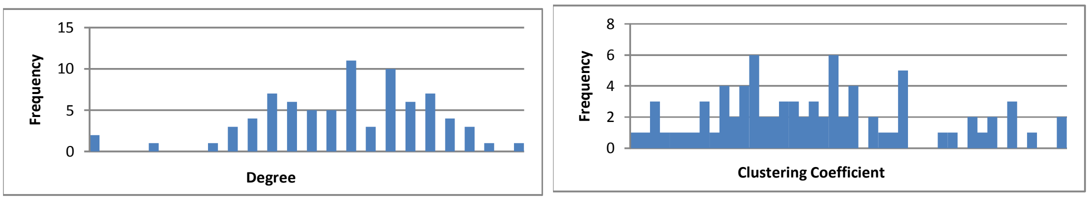
Table 4-5: Discovered Clusters for a 50 nodes topology using various input parameters

adv	***R_{thr}***	***No. of Clusters***	***Cluster Heads***	***Q (Modularity)***
1	∞	6	AP1(2), AP34 (10) , AP31(16), AP6(5), AP18(5), AP9(6)	0.500
1	**1**	4	AP34(11), AP31(21), AP18(7), AP9(7)	0.545
2	∞	6	AP1(2), AP34(11) , AP31(14), AP6(5), AP18(5), AP9(7)	0.542
2	**1**	5	AP34(11), AP31(14), AP6(5), AP18(8), AP9(7)	**0.572**
3	∞	5	AP34(11) , AP31(16), AP6(4), AP18(7), AP9(7)	0.552
3	**1**	4	AP34(11), AP31(21), AP18(7), AP9(7)	0.545

Figure 4-8-a presents the initial physical topology for a network graph of 80 APs which density is 0.186 and the average geodesic distance = 2.52. In this case we use a dense network graph without having obvious cluster structures (i.e., communities). Applying the proposed clusters formation scheme the higher Q value is achieved forming three clusters (Figure 4-8-c) and setting *adv=3 and* $R_{thr}=\infty$*.* Similar is the outcome for *adv=2 and* $R_{thr}=\infty$ *or* $R_{thr}=1.5$ (Table 4-6). We notice that the decrease of R_{thr} value some of the clusters are merged leading to a lower quality clusters (Q). This is due to the network density and the large number of inter-cluster edges.



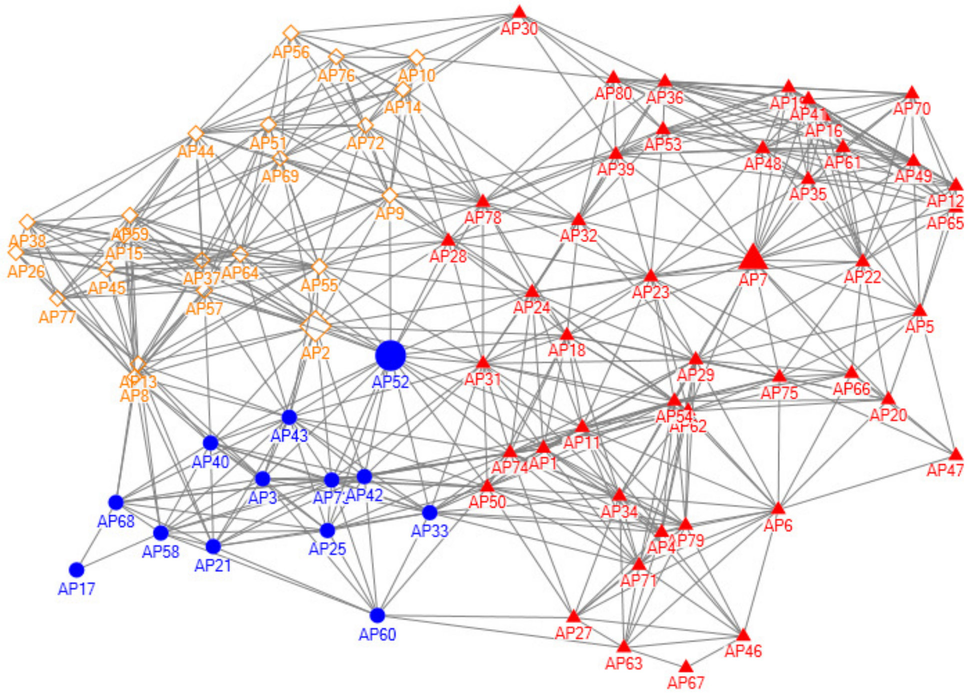
(a)



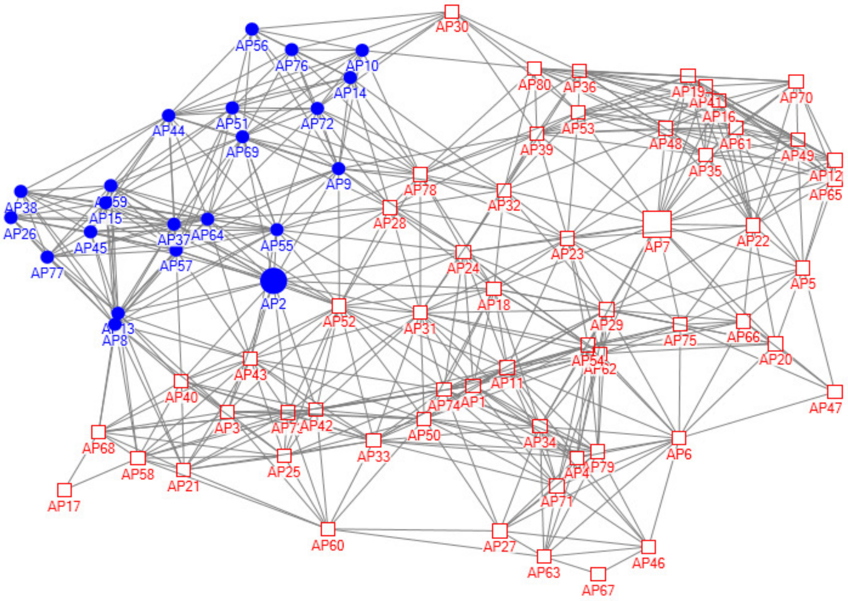
Minimum Degree	2
Maximum Degree	23
Average Degree	14.750

Minimum Clustering Coefficient	0.427
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.677

(b) **(c)**



(d)



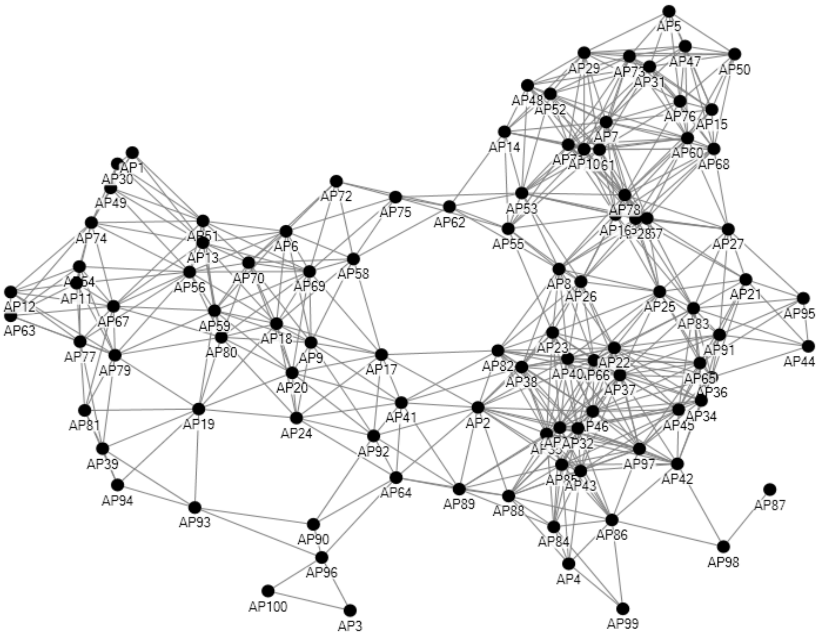
(e)

Figure 4-8: Topology of 80 Nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters visualization (adv=3, $R_{thr}=\infty$) (e) Clusters visualization (adv=3, $R_{thr}=1$)

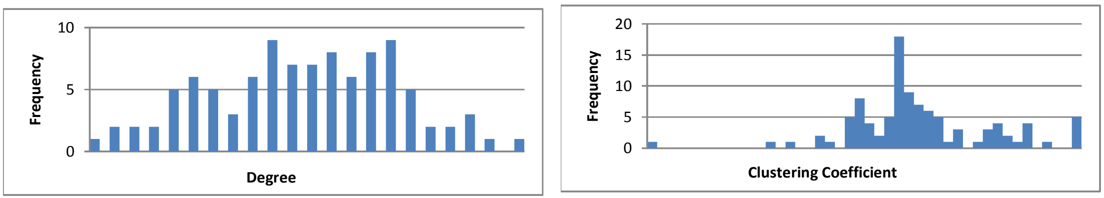
Table 4-6: Discovered Clusters for a 80 nodes topology using various input parameters

adv	*R_{thr}*	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	9	AP52(15), AP2(6), AP50(9), AP7(27), AP9(4), AP43(2), AP64(2), AP78(2), AP55(4)	0.289
1	**1**	2	AP52(48), AP7(30)	0.342
2	∞	5	AP52(12), AP2(12), AP50(4), AP7(38), AP9(9)	**0.413**
2	**1.5**	4	AP52(12), AP2(12) AP7(43), AP9(9)	**0.416**
2	**1**	2	AP2(22), AP7(56)	0.315
3	∞	3	AP52(12), AP2(21), AP7(44)	**0.419**
3	**1**	2	AP2(21), AP7(57)	0.307

Figure 4-9-a presents the initial physical topology for a network graph of 100 APs which density is 0.116 and the average geodesic distance = 3.56. Using the proposed clusters formation scheme three large clusters are discovered (AP56(33), AP37(37), AP7(24)) and a small one (AP96) that consists of only two members. Figure 4-9-b illustrates the constituent members as well as the intra-compartment edges of each cluster. Table 4-7 presents the clusters and the number of members that have been formed setting different values to *adv and R_{thr}* parameters.



(a)

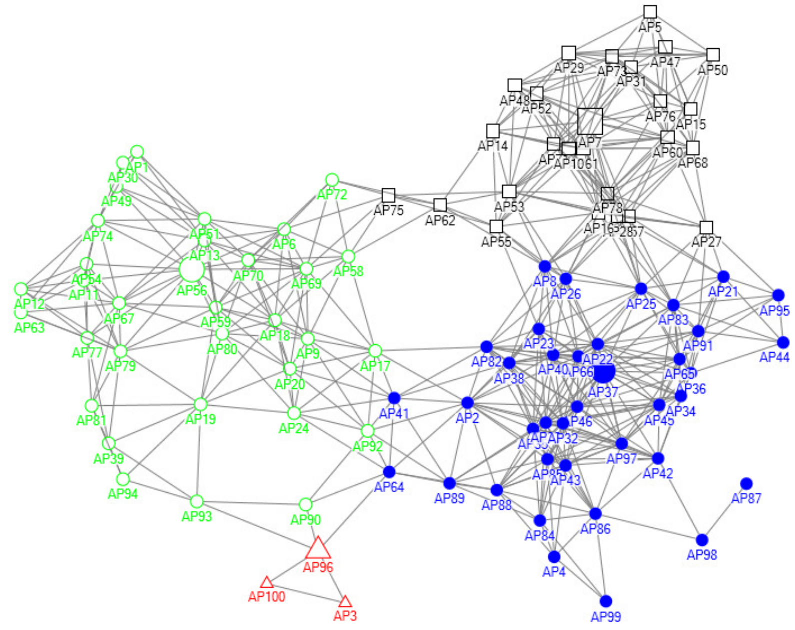


Minimum Degree	1
Maximum Degree	22
Average Degree	11.580

Minimum Clustering Coefficient	0.000
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.640

(b)

(c)



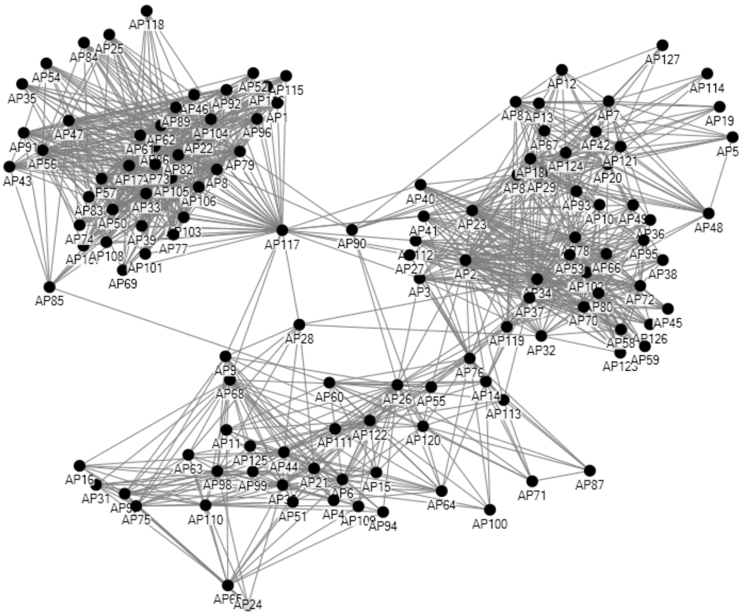
(d)

Figure 4-9: Topology of 100 Nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters Visualization (adv=3, $R_{thr}= 1$)

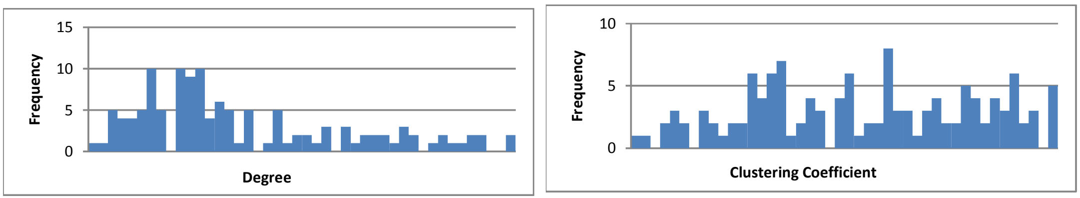
Table 4-7: Discovered Clusters for a 100 nodes topology using various input parameters

adv	R_{thr}	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	12	AP13(2), AP2(5), AP96(2), AP37(26), AP7(20), AP56(12), AP25(3), AP69(6), AP67(2), AP19(5), AP83(2), AP53(2)	0.448
1	**1**	3	AP56(31), AP37(43), AP7(23)	0.559
2	∞	8	AP2(2), AP96(2), AP37(30), AP7(24), AP56(25), AP69(2), AP19(4), AP83(3)	0.445
2	**1**	4	AP56(33), AP37(37), AP7(24), AP96(2)	**0.572**
3	∞	4	AP56(32), AP37(37), AP7(25), AP96(2)	**0.570**
3	**1**	4	AP56(32), AP37(37), AP7(25), AP96(2)	**0.570**

Figure 4-10-a presents the initial physical topology for a network graph of 127 APs which density is 0.142 and the average geodesic distance = 2.57. Using the proposed clusters formation scheme three large clusters are discovered (AP2(51), AP26(31), AP117(42)). Figure 4-10-b illustrates the constituent members as well as the intra-compartment edges of each cluster. Table 4-8 presents the clusters and the number of members that have been formed setting different values to *adv and* R_{thr} parameters.



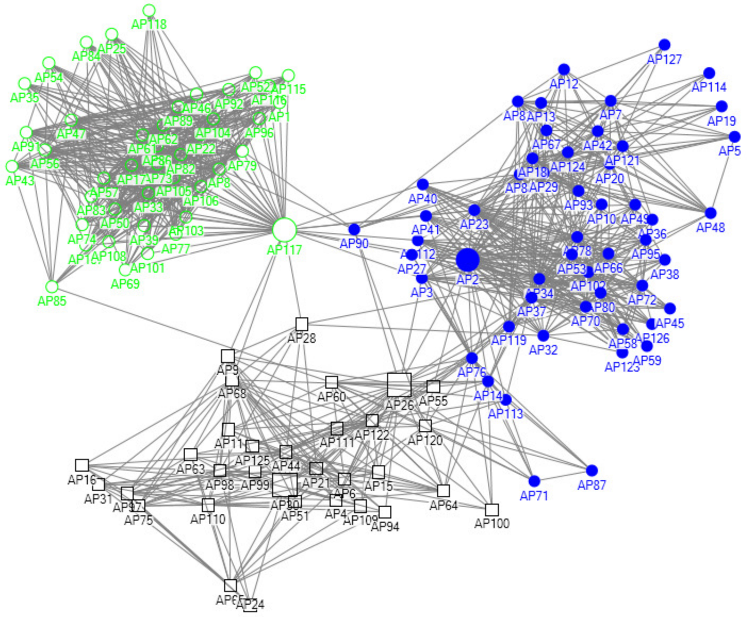
(a)



Minimum Degree	4
Maximum Degree	42
Average Degree	18.000

Minimum Clustering Coefficient	0.286
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.685

(b) **(c)**

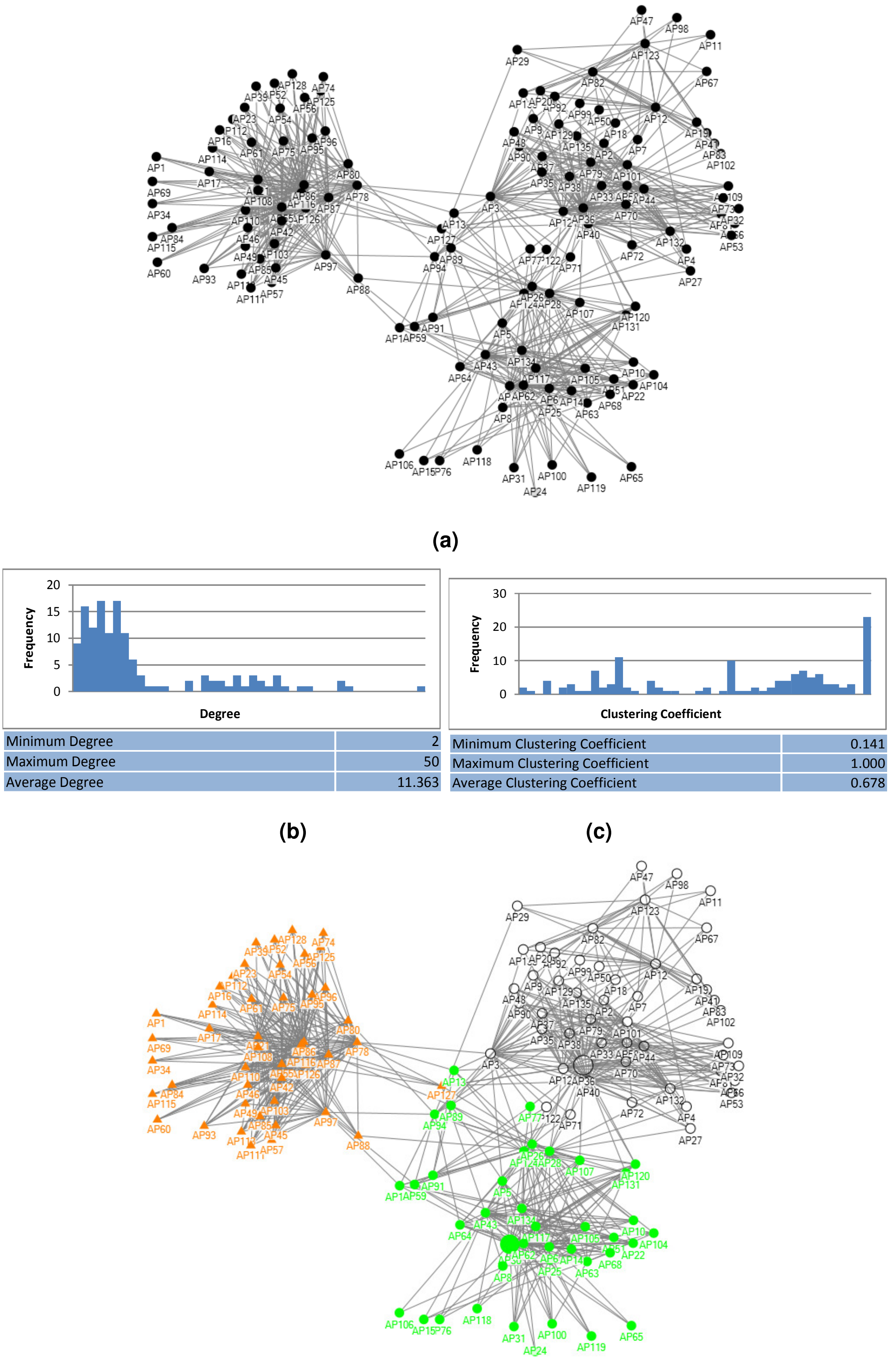


(d)

Figure 4-10: Topology of 127 nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters visualization (adv=3, R_{thr}= *1*)

Table 4-8: Discovered Clusters for a 127 nodes topology using various input parameters

adv	*R_{thr}*	*Number of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	7	AP117(35), AP2(39), AP26(19), AP30(9), AP88(7), AP20(2), AP82(9)	0.422
1	**1**	3	AP2(50), AP26(29), AP117(45)	0.554
2	∞	4	AP117(42), AP2(51), AP26(22), AP30(8)	0.560
2	**1**	3	AP2(51), AP26(31), AP117(42)	0.595
3	∞	3	AP117(42), AP2(51), AP26(31)	**0.595**
3	**1**	3	AP2(51), AP26(31), AP117(42)	**0.595**



Minimum Degree	2
Maximum Degree	50
Average Degree	11.363

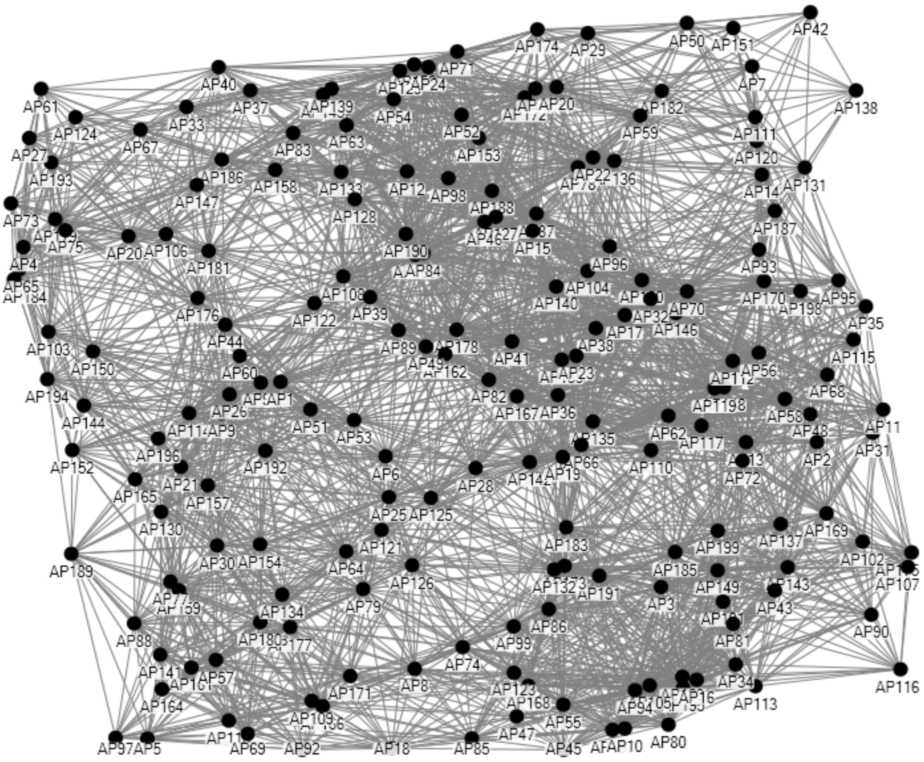
Minimum Clustering Coefficient	0.141
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.678

Figure 4-11: Topology of 135 Nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters visualization (adv=3, $R_{thr}= 1$)

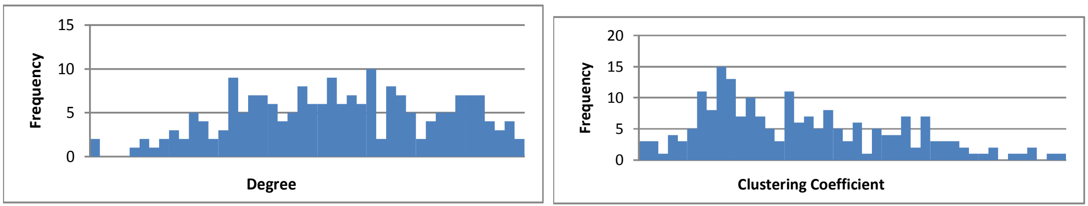
Table 4-9: Discovered Clusters for a 135 nodes topology using various input parameters

adv	*R_{thr}*	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	7	AP116(38), AP36(50), AP30(24), AP12(6), AP43(3), AP78(4), AP124(3).	0.428
1	**1**	3	AP116(43), AP36(61), AP30(28)	0.544
2	∞	4	AP116(44), AP36(49), AP30(39), AP12(15)	0.568
2	**1**	3	AP116(43), AP36(40), AP30(35), AP12(13)	0.602
3	∞	3	AP116(44), AP36(49), AP30(39)	**0611**
3	**1**	3	AP116(44), AP36(49), AP30(39)	**0.611**

Figure 4-11-a presents the initial physical topology for a network graph of 135 APs which density is 0.084 and the average geodesic distance = 2.78. This topology is similar to the previous one that consists of 127 nodes. Using the proposed clusters formation scheme three large clusters are discovered (AP116(44), AP36(49), AP30(39)). Figure 4-11-b illustrates the constituent members as well as the intra-compartment edges of each cluster. Table 4-9 presents the clusters and the number of members that have been formed setting different values to *adv and R_{thr}* parameters.



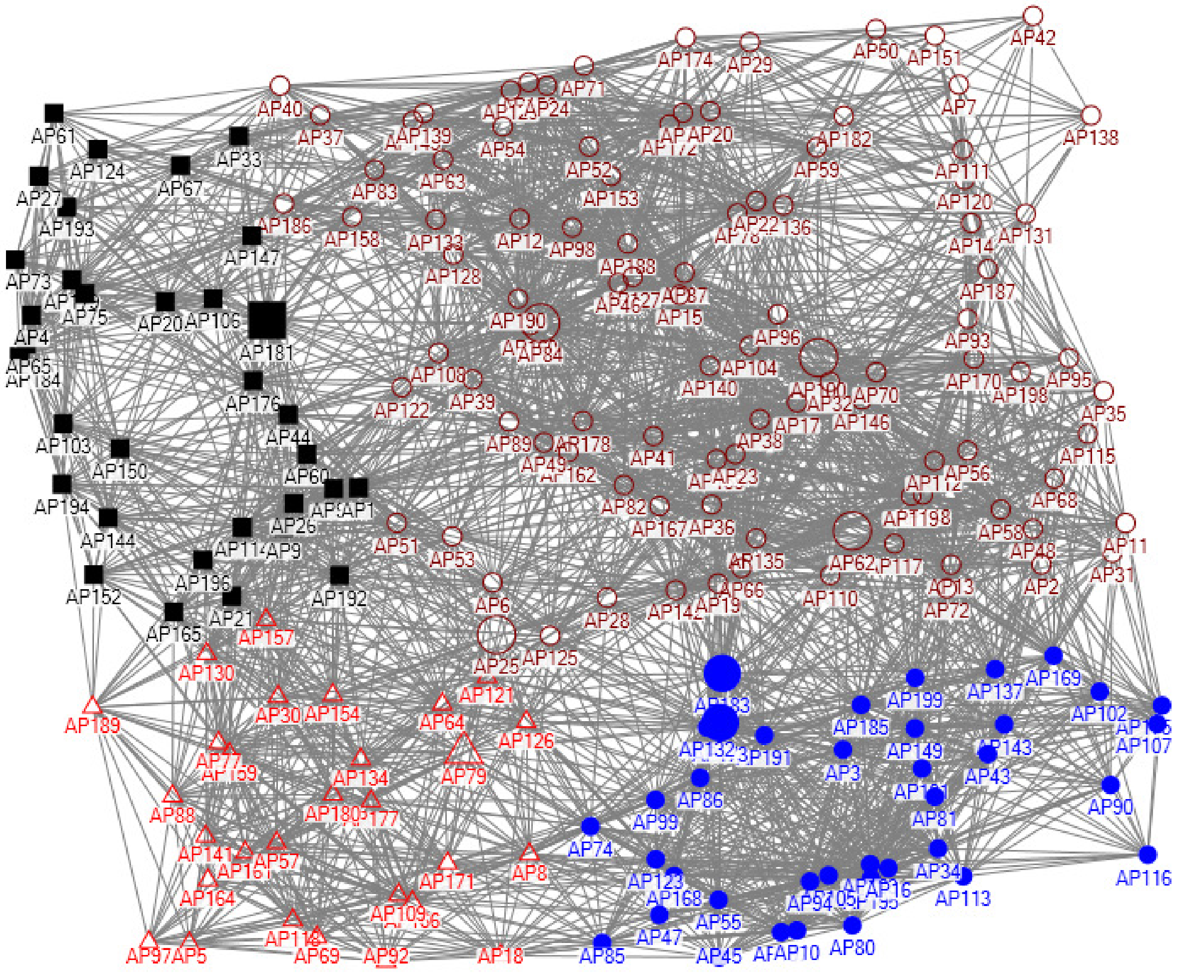
(a)



Minimum Degree	11
Maximum Degree	57
Average Degree	37.930

Minimum Clustering Coefficient	0.498
Maximum Clustering Coefficient	0.964
Average Clustering Coefficient	0.678

(b) (c)



(d)

Figure 4-12: Topology of 200 Nodes (a) Graph visualization, (b) Degree distribution, (c) Clustering Coefficient, (d) Clusters visualization (adv=2, $R_{thr}= 1.5$)

Figure 4-12-a presents the initial physical topology of a network graph that consists of 200 vertices (APs) and 3793 edges. Its density is 0.190 and the average geodesic distance = 2.44. In this dense graph the proposed clusters formation scheme creates four clusters (AP100(99), AP183 (37), AP79(28), AP181(32)), as it is illustrated in Figure 4-12-b. Similar to the topology that Figure 4-8presents the higher Q score is achieved for a high R_{thr} value (Table 4-10).

Table 4-10: Discovered Clusters for a 200 nodes topology using various input parameters

adv	R_{thr}	*No. of Clusters*	*Cluster Heads*	*Q (Modularity)*
1	∞	8	AP100(58), AP183 (33), AP79(29), AP181(45), AP62(3), AP59(16), AP173(5), AP13(3)	0.371
1	**1**	4	AP100(79), AP183 (43), AP79(29), AP181(45)	0.332
1	**1.5**	4	AP100(79), AP183 (43), AP79(29), AP181(45)	0.431
2	∞	4	AP100(99), AP183 (37), AP79(28), AP181(32)	**0.419**
2	**1**	3	AP100(132), AP183 (37), AP79(28)	0.332
2	**1.5**	4	AP100(99), AP183 (37), AP79(28), AP181(32)	**0.419**
3	∞	4	AP100(115), AP183 (28), AP79(35), AP181(18)	0.342
3	**1**	2	AP100(163), AP79(35)	0.172
3	**1.5**	4	AP100(115), AP183 (28), AP79(35), AP181(18)	0.342

The performed experiments presented above as well as the results of all tables show that the proposed algorithm discovers and forms efficiently the groups of a given network topology. The proposed algorithm could be implemented either in a distributed or in centralized manner collecting the network graph at a specific network location (e.g., network device, NMS). As expected the number of the formed clusters and consequently the number of the elected heads are reduced when the density level increases. However, in some of the experiments we notice that an increase of the density level might increase the number of elected heads (i.e., formed clusters), comparing to another less dense network graph. The reason for this fluctuation is that the increase of density (i.e., more overlaps among APs transmissions) is not allocated uniformly in the whole network area, but shows a specific locality. In some cases, even small changes in the number of edges (i.e., links among nodes) may lead to the creation of a new cluster. This effect justifies our design approach to avoid continuous re-organization of the formed clusters, since even small changes in the overlaps among APs may lead to large changes in the formed clusters.

In a uniformly dense graph (e.g., Table 4-6, Table 4-10) better clusters, in terms of the resulted modularity, are formed increasing R_{thr} parameter ($R_{thr}=1.5$, $R_{thr}=\infty$). On the other hand, in a sparse graph a larger number of advertisement hops (e.g., *adv=3*) and R_{thr}

value set at one ($R_{thr}=1$) facilitate the efficient discovery of group of nodes (e.g., Table 4-7, Table 4-8, Table 4-9). In the case that the groups of nodes could be easily identified (e.g., Table 4-8) then the set value of R_{thr} parameter does not affect significantly the performance of the proposed scheme.

CHAPTER 5

ENERGY SAVING AND OPTIMIZATION OF WIRELESS RESOURCES IN A SELF-ORGANIZING NETWORK

In SONs, the communication system is considered as an autonomic element, which is capable of monitoring its network-related state and modifying it based on policy rules that administrators have specified. The cognitive control loop (NECM, NDCM) is an enabler for the realization of in-network management that leads to its autonomy. As it is described in Chapter 3, the control loop includes processes for monitoring and perceiving a network node's internal state and environmental conditions, and then planning, deciding, and finally adapting, according to the identified conditions. The positions where control loops are placed, as well as their way of interacting, is a key design decision for the operation of network management and the performance of management functions (e.g., energy saving, CCO). The adoption of a totally centralized solution, on one hand, has the advantage that the management entity will have global view of the network status, which facilitates optimal decision making, but on the other hand there is a single point of failure, where scalability, computational, local optimization and local search issues arise. On the other hand, a fully distributed solution requires continuous coordination of control loop decisions, both spatially and temporally, in order to avoid conflicts or ripple effect phenomena.

Based on the algorithm described in Chapter 4, a hybrid approach has been adopted, where APs form clusters in order to perform their management functions: a) coverage and capacity self-optimization, and b) energy management. The control loop of the head, which is unique per cluster, periodically collects and correlates monitoring data provided by simple member nodes of the corresponding cluster. Based on this information the head builds a greater view of the status of the cluster area, and thus has the capability to solve performance or configuration problems of the cluster area that each simple member cannot handle individually. The head of each cluster undertakes energy efficiency together with CCO of the region specified by the cluster borders.

5.1 Mapping of System Capabilities to a WiFi Radio Technology

The head AP retrieves topological, performance, and energy consumption information from the member APs. Moreover, the head receives monitoring data that the associated UEs provide to the respective AP. All these data allow the head to build its situation awareness, which includes energy and CCO awareness of the cluster. Energy awareness describes the energy consumption of the network area, and of specific nodes (AP and UE). CCO awareness provides the head with information that is related to the available capacity, the data traffic, and the coverage overlap of the APs of the cluster.

Energy and CCO awareness are the input for the decision-making phase, which consists of two steps. In the first step, the head uses a rule-based scheme in order to evaluate the existing situation awareness and select one of the following types of configuration action:

- AP dynamic deactivation (reactivation) and load balancing,
- UE multi-hop relay formation,
- AP channel re-selection.

Thresholds and rules set by the network administrator guide the triggering of the appropriate adaptation for energy saving and CCO management. In the second step, the decision making engine resolves the deduced configuration action type. Finally, the head transmits the decision making output to the underlying APs, which proceed to the enforcement of the adaptation. Figure 5-1 depicts the main functions of the head AP.

At this point, we should clarify that energy awareness facilitates energy saving actions, while CCO awareness activates the corresponding configuration actions for optimizing the provided bandwidth or coverage, according to the end user traffic levels. However, as it is analyzed below, these actions are not independent, owing to the dependency of the affected metrics. A CCO action might increase energy consumption for one type of device, or it might subsequently require a complementary action for energy saving.

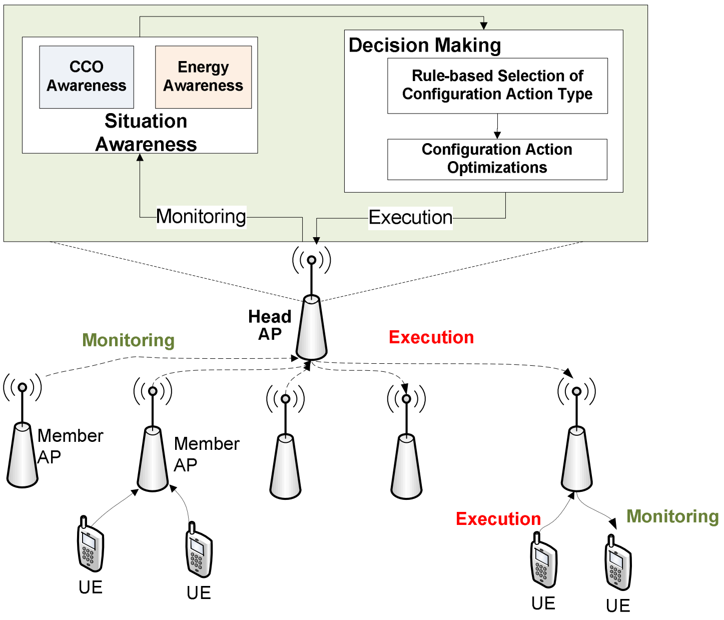


Figure 5-1: Cluster head management tasks

Similarly, an energy saving action might lead to the degradation of a CCO metric. The decision making of the head takes into account the above-mentioned dependencies, and addresses energy efficiency jointly with other management tasks [136], [137], [138].

5.2 Building Situation Awareness

Situation awareness is the step that prepares the decision-making phase. The periodic updates that APs provide to the head of the cluster allow the latter to calculate the CCO and energy specific metrics.

5.2.1 Coverage and Capacity Levels

Coverage and capacity management consists of two different but interrelated tasks, especially for wireless LANs, which are inherently more complex than other wireless network technologies, due to the use of RF links and the unplanned (or in many cases random) placement of APs. Each AP configuration uses different RF characteristics, and the dynamic nature of the wireless channel makes the communication environment more volatile. The goal of coverage management is to provide network connectivity at all desired locations, while capacity management undertakes to provide sufficient bandwidth to satisfy clients' communication needs.

The number of deployed APs as well as the overlappings of their transmission ranges are taken into account. We assume that a cluster is represented by a connected and undirected graph $G = (V, E)$, where V is the set of APs in the cluster and E is the set of edges (connectivity links) between APs. An edge exists between two APs, if one is within the coverage area of the other. Let the cardinalities of V and E be denoted by n, and e, respectively; i.e., n is the number of APs that constitute the cluster, while e is the number of the existing connectivity links among the APs of the cluster. For the calculation of the degree of coverage overlap in a cluster we use the clustering coefficient (*CC*), used in graph theory [111]:

$$CC = \frac{2e}{n(n-1)} \tag{5-1}$$

A high *CC* value indicates that there are many coverage overlaps among the available APs. However, *CC* gives little information if n is small. For this reason, we introduce a smoothing parameter s in order to scale *CC* according to the actual size of the cluster, and define a modified *CC*, referred to as the overlapping factor (*OF*), as follows:

$$OF = s \cdot CC \tag{5-2}$$

where

$$s = \begin{cases} \frac{n}{a}, & n < a \\ 1, & n >= a \end{cases} \tag{5-3}$$

and *a* is a predefined positive integer threshold value. As can be seen from (5-3), for good size clusters, the default value of the smoothing parameter *s* is unity; however, for cluster sizes below the threshold, the confidence for the coverage overlap and the avoidance of coverage holes is reduced proportionally with the cluster size. The Capacity Usage Ratio (*CUR*) of a cluster network area with *n* APs is defined as the fraction of the available capacity that is actually being used:

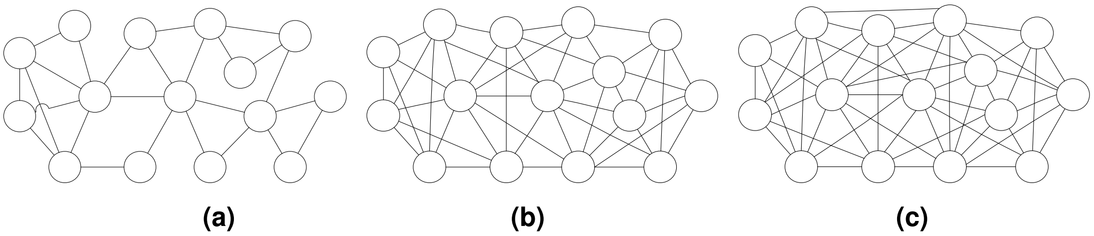
$$CUR = \frac{\sum_{i=1}^{n} C_i}{\sum_{i=1}^{n} C_i^{\max}} \tag{5-4}$$

where C_i and $C_i^{\max}$ is the used (uplink and downlink) capacity and the maximum available (uplink and downlink) capacity, respectively, of AP *i*.

The correlation of the *CUR* with the *OF* of the APs in a cluster area allows for more effective interpretation of the information that CUR provides, by taking into account the overlap level of the offered bandwidth. For this reason we use the composite metric of Coverage Optimization Opportunity (COOP) introduced in [111]:

$$COOP = CUR^{OF} \tag{5-5}$$

The *COOP* metric is useful for the identification of optimization opportunities for low load situations, where less capacity needed, as well for high load situations, where more capacity is required. A low COOP value means that too much capacity is provided in a very dense area, while a too high *COOP* value indicates an overloaded network area, where more resources are needed. Figure 5-2 shows the calculated COOP value of three clusters, each with *15* APs, but with different levels of density (*OF* values) and CUR values.



(a) **(b)** **(c)**

Figure 5-2: COOP values for a topology of n=15 APs (a) OF= 0.24, CUR= 0.1, (b) OF= 0.45, CUR= 0.1, (c) OF= 0.51, CUR= 0.6

5.2.2 Energy Consumption

For efficient energy management, it is important to analyze energy consumption of a wireless LAN at the device, component and function levels. The total energy consumption (EC_{CL}) of a cluster that consists of n APs and m UEs is defined as follows:

$$EC_{CL} = \sum_{j=1}^{m} EC_{UE_j} + \sum_{i=1}^{n} EC_{AP_i} \tag{5-6}$$

where EC_{AP_i} denotes the energy consumption of AP_i and EC_{UE_j} denotes the energy consumption of UE_j.

Regardless of the type of the network node, the energy consumption could be estimated for the processing phase and the communication phase of a data packet. Hence, the Energy Consumption (*EC)* for a link among UEs and other network nodes is calculated when the respective node transmits or receives a data packet. Adopting the methodology of [139], for a link among two nodes, *EC* includes the dissipated energy for the transmission (EC_{Tx}) and for the reception of a data packet (EC_{Rx}). Energy consumption, during the processing phase (EC_{pr}) [140] is not estimated in this article:

$$EC = EC_{Tx} + EC_{Rx} + EC_{pr} \tag{5-7}$$

More specifically, the transmitter consumes energy to run the radio electronics and the power amplifier:

$$EC_{Tx} = EC_{Tx,electronics} + EC_{Tx,amplifier} \tag{5-8}$$

$$EC_{Tx,electronics} = lE_{elec} \tag{5-9}$$

$$EC_{Tx,amplifier} = lE_{fs}d^2 \tag{5-10}$$

Equation (9) is used for a free space channel model, while $EEC_{Tx,amplifier} = lE_{mp}d^4$ for a multipath fading channel model, where *d* is the distance among the transmitter and the receiver, and *l* is the number of bits in a data packet. EC_{elec} denotes the energy consumed to run the radio electronics for one bit (Joules/bit), while E_{fs} (Joules/bit/m^2) and E_{mp} (Joules/bit/m^4) denote the consumed energy for the power amplifier. The last three parameters are constant and RAT specific.

On the other hand, the receiver consumes energy only to run the electronics $EC_{Rx} = lE_{elec}$. According to Equation (10), the distance (*d*) between the transmitter and the receiver is the main parameter that could affect directly the energy consumption, optimizing the topology and the active links of a network area.

Furthermore, the Bit Error Rate (BER) is a second parameter that indirectly affects energy consumption, since it leads to data packet retransmissions. If the BER has been increased (e.g., after a topology change) the consumed energy of the transmitter (and the receiver) is increased for no purpose, since more transmissions will take place for the same volume of data.

Following the above analysis, we could extend formula (7) by analyzing further energy consumption of each AP *i* and end UE *j*, during the transmission and reception phases:

$$EC_{AP_i} = EC_{AP_i}^{Tx} + EC_{AP_i}^{Rx} \tag{5-11}$$

$$EC_{UE_j} = EC_{UE_j}^{Tx} + EC_{UE_j}^{Rx} \tag{5-12}$$

5.3 Rule-based Selection of Configuration Types

The periodic measurements that APs send to the head, and the situation awareness that the head builds, calculating CCO and energy-related metrics, are the inputs to the decision making phase. In this phase, the head selects the appropriate configuration action type for energy saving and then decides on the specific configuration action to be used. Finally, the head sends its decision to the member APs to enforce the reconfiguration. The adaptations that we use for energy saving are the following:

- *AP dynamic deactivation* reduces radio resources wasting and network side energy consumption, taking into account the existing coverage and capacity needs. The deactivation of an AP is followed by a load balancing scheme, which undertakes the hand-over of participating UEs. APs reactivation is prioritized when there is need for more resources in the cluster area (section 5.4.1).
- *UE multi-hop relay formation* deploys hybrid wireless networks using multi-hop and homogeneous RAT links. The goal is the reduction of energy consumption at the UE side that derives from the data packet transmission phase. The trigger for UE multi-hop relay formation comes from the head, and the UEs undertake to form the appropriate local topology (section 5.4.2).
- *AP channel re-selection* changes the selected frequency channel to reduce the overlaps among neighboring cells. Consequently, it reduces UE and AP energy consumption for the reception of data traffic that comes from neighboring cells (section 5.4.3)

The energy behavior of a device, a component or a function depends on that of others, since their parameterization or re-configuration of an entity affects the energy consumption of other entities. An energy saving action for a specific network node might substantially modify the network topology and configuration, thus, affecting the energy consumption of other devices or components. For instance, an adaptation that aims at the rational usage of radio resources (e.g., AP deactivation) will decrease energy consumption for a specific device type (i.e., AP) and function (e.g., Rx, Processing). On the other hand, it might increase the consumed energy for UE transmissions. Consequently, it is necessary to measure the changes in energy consumption after each adaptation and between monitoring periods. To this end, we introduce below the corresponding metrics for energy consumption change, which will assist the decision making for network adaptation.

Let ΔEC_{UE} denote the change of energy consumption of the UE side between two monitoring periods (*t, t-1),* measured in Joules/bit. ΔEC_{UE} depends on the data traffic received and transmitted by the involved nodes and is calculated as follows:

$$\Delta EC_{UE} = \left(\frac{\sum_{\mathrm{j=1}}^{m} EC_{UE_j,t}}{\sum_{\mathrm{j=1}}^{m} (T_{UE_j,t}^{Tx} + T_{UE_j,t}^{Rx})} - \frac{\sum_{\mathrm{j=1}}^{m} EC_{UE_j,t-1}}{\sum_{\mathrm{j=1}}^{m} (T_{UE_j,t-1}^{Tx} + T_{UE_j,t-1}^{Rx})} \right) \tag{5-13}$$

In (5-13) $EC_{UE,t}$ and $EC_{UE,t-1}$ denote the energy consumption (in Joules) of all UEs that participate in the cluster, during the monitoring period *t* and the previous one (*t-1*), respectively. $T_{UE_j,t}^{Tx}$ denotes the transmitted traffic (in bits) and $T_{UE_j,t}^{Rx}$ the received data traffic (in bits) of UE *j*, during the monitoring period *t*. The change in energy consumption is allocated between the reception and the transmission phase.

Let the metric ΔEC_{UE}^{Tx} denote the cluster-level change of the consumed energy (in Joules/bit) for UE data transmission; it is calculated as follows:

$$\Delta EC_{UE}^{Tx} = \left(\frac{\sum_{\mathrm{j=1}}^{m} EC_{UE_j,t}^{Tx}}{\sum_{\mathrm{j=1}}^{m} T_{UE_j,t}^{Tx}} - \frac{\sum_{\mathrm{j=1}}^{m} EC_{UE_j,t-1}^{Tx}}{\sum_{\mathrm{j=1}}^{m} T_{UE_j,t-1}^{Tx}} \right) \tag{5-14}$$

Next, let the metric ΔEC_{UE}^{Rx} denote the cluster-level change of the consumed energy (in Joules/bit) for UEs reception of data traffic that is transmitted by neighboring UEs ($T_{UE_j}^{Tx}$) and APs ($T_{AP_i}^{Tx}$). ΔEC_{UE}^{Rx} is given as follows:

$$\Delta EC_{UE}^{Rx} = \left(\frac{\sum_{j=1}^{m} EC_{UE_j,t}^{Rx}}{\sum_{j=1}^{m} T_{UE_j,t}^{Tx} + \sum_{i=1}^{n} T_{AP_i,t}^{Tx}} - \frac{\sum_{j=1}^{m} EC_{UE_j,t-1}^{Rx}}{\sum_{j=1}^{m} T_{UE_j,t-1}^{Tx} + \sum_{i=1}^{n} T_{AP_i,t-1}^{Tx}} \right) \tag{5-15}$$

where $T_{AP_i,t-1}^{Tx}$ is the transmitted data traffic (in bits) of AP *i*, during the monitoring period *t-1.*

Similarly, the change in energy consumption at the AP side, denoted by *ΔEC_{AP}*, for the two monitoring periods, takes into account the consumed energy of all APs in correlation with the data traffic that APs transmit($T_{AP_i}^{Tx}$) and receive($T_{AP_i,t}^{Rx}$):

$$\Delta EC_{AP} = \left(\frac{\sum_{i=1}^{n} EC_{AP_i,t}}{\sum_{i=1}^{n} (T_{AP_i,t}^{Tx} + T_{AP_i,t}^{Rx})} - \frac{\sum_{i=1}^{n} EC_{AP_i,t-1}}{\sum_{i=1}^{n} (T_{AP_i,t-1}^{Tx} + T_{AP_i,t-1}^{Rx})} \right) \tag{5-16}$$

We use ΔEC_{AP}^{Tx} to denote the change in energy consumption for the data transmission of all APs that constitute the cluster:

$$\Delta EC_{AP}^{Tx} = \left(\frac{\sum_{i=1}^{n} EC_{AP_i,t}^{Tx}}{\sum_{i=1}^{n} T_{AP_i,t}^{Tx}} - \frac{\sum_{i=1}^{n} EC_{AP_i,t-1}^{Tx}}{\sum_{i=1}^{n} T_{AP_i,t-1}^{Tx}} \right) \tag{5-17}$$

Let ΔEC_{AP}^{Rx} denote the change in energy consumption for APs reception of data packets:

$$\Delta EC_{AP}^{Rx} = \left(\frac{\sum_{i=1}^{n} EC_{AP_i,t}^{Rx}}{\sum_{j=1}^{m} T_{UE_j,t}^{Tx} + \sum_{i=1}^{n} T_{AP_i,t}^{Tx}} - \frac{\sum_{i=1}^{n} EC_{AP_i,t-1}^{Rx}}{\sum_{j=1}^{m} T_{UE_j,t-1}^{Tx} + \sum_{i=1}^{n} T_{AP_i,t-1}^{Tx}} \right) \tag{5-18}$$

The proposed scheme for the selection of the configuration action type is described in Table 5-1. The head considers coverage and capacity levels as well as the changes in the energy consumption of devices (UE, AP) and functions in order to reach a more energy efficient state for the cluster area.

The head AP firstly checks whether the deactivation or the reactivation of one or more APs that exist in a cluster area is necessary so as to address low and high *COOP* values, respectively. APs switch-off procedure is triggered if

$$COOP < B_d \tag{5-19}$$

where B_d is the threshold for APs deactivation. The process for APs switch-on is triggered if

$$COOP > B_r \tag{5-20}$$

Table 5-1: Scheme for cluster head rule-based decision making

```
Cluster Head: Select Configuration Action Type
Calculate COOP;
$CA_t$=0;
If COOP < $B_d$
    AP_deActivation();
    LoadBalancing();
    $CA_t$=1;
else If COOP > $B_r$
    AP_reActivation();
    LoadBalancing();
    $CA_t$=1;
else If $B_d$ ≤ COOP ≤ $B_r$ or $CA_{t-1}$ == 1
    Calculate $\Delta EC_{UE}$, $\Delta EC_{UE}^{Tx}$, $\Delta EC_{UE}^{Rx}$, and $\Delta EC_{AP}^{Rx}$
    If $\Delta EC_{UE}^{Rx} > B_e$ or $\Delta EC_{AP}^{Rx} > B_e$
        ChannelReallocation();
    else if $\Delta EC_{UE}^{Tx} > B_e$
        BuildmultiHopPath();
    End
End
```

where B_r is the threshold for APs reactivation. The thresholds B_d and B_r are set by the network administrator. An optimization opportunity for a low load situation indicates that there is the possibility to deactivate one or more APs. The goal is to avoid wasting resources in the cluster area, without concurrently reducing the appropriate geographical coverage of the APs. For that reason a high *OF* value in conjunction with a low load situation (low *CUR*) is required. Similarly, in a high load situation, the head estimates the necessity to reactivate an AP in order to address the increased capacity needs.

After the enforcement of an AP deactivation/reactivation the topology of the corresponding network area changes, which in turn might affect negatively the energy consumption of other functions, or even neighboring devices (UE, AP). Hence, if there is no need for APs deactivations/reactivations ($B_d \leq COOP \leq B_r$), or if there is no available AP to switch-On/Off ($CA_{t-1} = 1$), the head should assess the change in energy consumption of other devices and components in order to decide on further adaptations.

Specifically, it checks whether the energy consumption in the UE side has been increased beyond a certain threshold, B_e i.e., it checks whether $\Delta EC_{UE} > B_e$. The value of the threshold B_e (measured in Joules/bit) is set to reflect a substantial increase of the energy consumption in the cluster area. If the energy consumption of UEs has been increased mainly for the transmission phase (ΔEC_{UE}^{Tx}), then the creation of a shorter (multi-hop) link is investigated by each *UE* that shows an increase of the consumed energy. Moreover, if there is an increase of the consumed energy for the reception phase, either in the UE side (ΔEC_{UE}^{Rx}) or in the AP side (ΔEC_{AP}^{Rx}), then the channel re-allocation action is selected.

5.4 New Algorithms for Energy Efficiency and Coverage and Capacity Optimization

In this section, we describe the algorithms for energy efficiency and CCO in the context of SONs. As mentioned in previous sections, each AP that participates in the cluster monitors its status (energy, traffic) and discovers its local physical topology (e.g., neighboring APs). In addition, AP_i requests the associated UEs to provide their list of sensed nodes. AP_i collects periodically the above data, which is transmitted to the head of the cluster together with other information about its operational status (e.g., EC_{AP_i}, EC_{UE_j}, $T_{AP_i}^{Tx}$, $T_{AP_i}^{Rx}$, $T_{UE_j}^{Tx}$, $T_{UE_j}^{Rx}$, C_i, $C_i^{\max}$). Hence, the head based on the collected information updates its awareness by building the cluster-level topology graph, which includes the following data:

- A_{AP_i} : set of APs that are sensed by AP_i
- A_{UE_j} : set of APs that are sensed by UE_j
- $\bar{A}_{UE_j}$: the AP where UE_j is associated with
- $\bar{U}_{AP_i}$: set of UEs that are associated with AP_i
- U_{UE_j} : set of UEs sensed by UE_j
- $\bar{U}_{UE_j}$: set of UEs sensed by UE_j that are associated with the same AP ($\bar{A}_{UE_j}$)

5.4.1 Dynamic Access Point Deactivation/Reactivation and Load Balancing

The dynamic deactivation or reactivation of one or more APs is used to achieve a more efficient usage of radio and energy resources at the network side. In a specific geographical area, the available wireless resources (i.e., APs) might be underutilized for a long period of time juxtaposed with the capacity requirements (e.g., throughput, number of users). Thus, in terms of radio and energy resources, the deactivation of a set

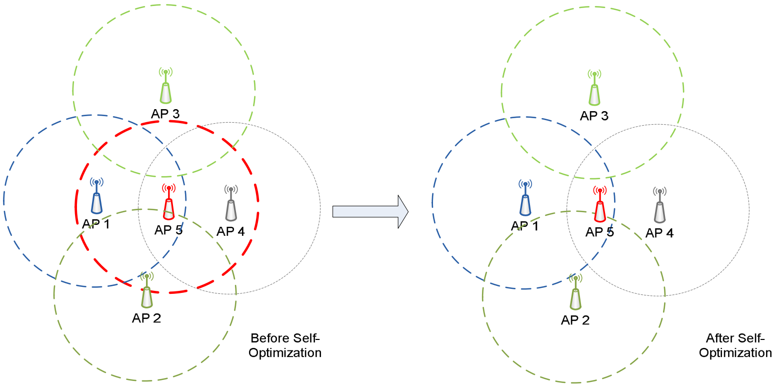


Figure 5-3: Sample topology before and after self-Optimization

of APs could be beneficial for network areas that have more APs than actually needed, with possible reactivation when the network conditions necessitate more capacity. An AP deactivation means that the AP is set to a standby/idle mode, where it does not transmit beacon packets, and UE is not able to detect its presence. The latter is achieved through the deactivation of specific wireless interface capabilities (RF, baseband). The scheme for the selection of the appropriate AP to switch-on or switch-off, and the corresponding reallocation of UEs are described below.

The head AP calculates the *OF* and the *CUR* metrics, using equations (5-2) and (5-4), respectively, and then the cluster area *COOP* value. If the *COOP* of the network cluster area has reached the level that satisfies (5-19), the AP switch-off action is triggered (Table 5-2). Firstly, the head AP builds the list of candidate APs for deactivation. This list includes those APs for which all their associated UEs ($UE_j \in \bar{U}_{AP_i}$) have the capability to hand over to a neighboring AP ($AP \in A_{UE_j}$), satisfying their $C_{i,j}$ requirements. Then, the head calculates the local COOP value of each candidate AP_k, as follows:

$$COOP_{AP_k} = \left(\frac{C_k}{C_k^{\max}} \right)^{OF_k}$$

where OF_k is the overlapping factor of the sub-graph that is formed by AP_k and its one-hop away APs; equation (5-2) is used for the OF_k calculation. The head selects for deactivation the access point AP_k^* for which $COOP_{AP_k}$ is minimum.

After the selection of the appropriate AP for deactivation, the head of the cluster proceeds to the reallocation of the UEs that are associated with AP_k^* ($UE \in \bar{U}_{AP_k^*}$). The head of the cluster, firstly, prioritizes the *UEs* of AP_k for the handover process, and then selects the AP where each UE should be handed over. The head starts with those UEs that are closer to AP_k in order to avoid the increase of $EC_{UE_j}^{Tx}$ after the re-allocation, subsidizing also those UEs that have a small number of neighboring APs. To facilitate this procedure, we introduce the parameter β_{UE_j}:

$$\beta_{UE_j} = w_1\left(1-\frac{d_{k,j}}{d_k^{\max}}\right) + (1-w_1)\left(1-\frac{v}{n}\right) \tag{5-21}$$

where $d_{k,j}$ is the distance between UE_j and *AP*$_k$, $d_k^{\max}$ is the transmission range of *AP*$_k$, *v* denotes the number of APs that UE_j senses (i.e., A_{UE_j}) and *n* the number of APs that constitute the cluster. The weight w_1, $0 \le w_1 \le 1$, is set by the system administrator in accordance with the importance of each of the two terms in (5-21).

The head of the cluster selects the UE that has the maximum β_{UE_j} value and searches for $AP_i \in A_{UE_j}$ where UE_j should be handed over. UE_j selects the AP_i ($AP_i \in A_{UE_j}$) that has the minimum γ_{AP_i} value:

$$\gamma_{AP_i} = w_2\left(1-\frac{d_{k,i}}{d_i^{\max}}\right) + (1-w_2)\left(1-\frac{CUR}{C_i}\right) \tag{5-22}$$

Through equation (5-22) it is assured that each UE is allocated to the closest distance AP, without overloading a neighboring AP, taking into account the CUR of the cluster. The weight w_2, $0 \le w_2 \le 1$, is set by the system administrator to reflect the importance of each of the two terms in (5-22).

In the case that the cluster-level *COOP* value satisfies inequality (5-20), then the process for AP reactivation is initiated. The head firstly checks for deactivated APs that could be enabled in order to serve the increased capacity requirements. If more than one APs are available for reactivation, then the head selects AP_r', which has the maximum local *COOP* ratio, which is calculated as follows:

$$COOP_{AP_r} = \left(\sum_{i=1}^{z} \frac{C_i}{C_i^{\max}}\right)^{OF_r},$$

where *z* denotes the number of one-hop away neighbors of *AP*$_r$, *OF*$_r$ is the overlapping factor of the sub-graph that candidate *AP*$_r$ and its *z* neighboring APs form. The goal is to find a high load area with a small overlapping factor.

After the reactivation of AP_r', the head builds the list of UEs that can sense the newly activated AP. The scheme selects to handover the UEs that are associated with an *AP*$_i$

Table 5-2: Algorithm for AP deactivation

```
Cluster Head: Dynamic AP deactivation and load balancing scheme

$G \leftarrow$ List of candidate APs for deactivation;
For each $AP_k \in G$
    Calculate $COOP_{AP_k}$;
End
$AP_k^* \leftarrow$ Select $AP_k$ with minimum $COOP_{AP_k}$ to deactivate;
For each $UE_u \in \bar{U}_{AP_k^*}$
    Calculate $\beta_{UE_u}$;
End
Do
    Select UE with maximum $\beta_{UE_u}$;
    For each $AP_i \in A_{UE_u}$
        Calculate $\gamma_i$;
    End
    Select AP with minimum $\gamma_i$ where $UE_u$ should be associated;
    Remove $UE_u$ from $\bar{U}_{AP_k^*}$
while $\bar{U}_{AP_k^*} \neq \varnothing$
```

that has CUR_i higher than the *CUR* of the cluster and starts with the UE that is closer to AP_r'. The process stops when the CUR of AP_r' exceeds the CUR of the cluster.

5.4.2 User Equipment Multi-Hop Relay Formation

In this chapter, we present the scheme for the formation of multi-hop relays among UEs. The goal is to exploit local networking capabilities and thus reduce the energy consumed by UEs for the transmission of their UL traffic. An existing (one-hop) data link, which has been established between a UE and an AP, could be replaced by a new multi-hop path that is less energy demanding, leading to the same or different AP. In the sequel, we will use the following definitions:

- Current Energy Consumption (CEC): is the consumed energy of UE's currently active data link for the transmission of data packets, measured in Joules/bit.

- Current Data Rate (CDR): is the data rate of UE's UL, measured in bps.
- Expected Data Rate (EDR): has been introduced for wireless ad-hoc networks in [141], and provides an estimation for the data rate between a UE and the next hop nodes (UE, AP).
- Expected Energy Consumption (EEC): has been inspired by the EDR metric and provides an estimation for the expected energy consumption of a data packet between a UE and the next hop nodes (e.g., UE, AP), focusing on the transmission phase. The estimated distance d (in meters) among the involved nodes and the BER provide an indication for EEC, by extending (5-8) as follows:

$$EEC(d, PER) = EEC_{Tx} + BER \cdot EEC_{Tx} \qquad (5\text{-}23)$$

Let UE_0, be the node that we are attempting to reduce its *CEC*. UE_0 is also the root of the multi-hop path, which will terminate at an AP. The head of the cluster controls the topology re-organization collecting local information from UE_0 and its neighboring nodes (UE, AP). UEs calculate the *EDR* and the *EEC* for the each link that constitutes a potential data path (e.g., Figure 5-4). The goal is to discover and form a multi-hop path that minimizes UE_0 energy consumption subject to *CDR* constraint and without increasing the energy consumption of the entire path (i.e., end-to-end) (EEC_{e2e}) higher than the *CEC* ($EEC_{e2e} < CEC$).

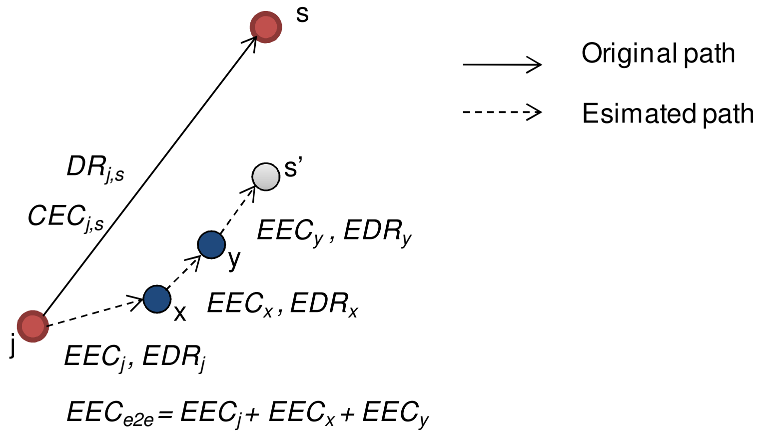


Figure 5-4: Sample topology for UEs multi-hop relay

The solution proposed below provides the means for the development of the appropriate multi-hop link, without the need to explore all possible networking solutions. In a dense network environment, with hundreds of nodes (e.g., UEs, smart objects), the building of the appropriate multi-hop path is a demanding task both in terms of the communications and the computations required. The search space includes all possible multi-hop paths that could be formed from the initiating UE_0, while the solution space includes those multi-hop paths that satisfy the necessary energy or QoS criteria. The search space is a superset of the solutions space. The root of all multi-hop paths of the solutions space is UE_0, while all links terminate at an AP. The intermediate nodes along the path are UEs (e.g., Figure 5-4). The formation of the multi-hop path involves three main phases:

1. The head of the cluster defines the search space in terms of hops from the initiating node UE_0 denoted by H_{stop}. The head sends H_{stop} to UE_0.
2. Then the head AP requests from the UEs that are in the search space to discover neighboring UEs and APs and to calculate the *EEC* and *EDR* for each one hop away node (UE, AP) that has the available resources to collaborate (e.g., battery level). The creation of the local solutions space starts from UE_0 (i.e., the root of the multi-hop path) and it is extended, based on a recursive procedure that is described in Table 5-3. Each UE builds its local solutions space and sends to the head of the cluster the identified paths towards an AP.
3. The head collects UEs notifications and composes the total solutions space for UE_0. Finally, the head AP selects the most energy efficient path, solving the following problem:

$$\begin{aligned} &Minimize \quad EEC_j(d_j, BER_j) \\ &s.t.\ EDR_j \geq CDR_{j,s} \\ &P(n)\sum_e EEC_e(d_e, BER_e) < CEC_{j,s} \end{aligned} \tag{5-24}$$

d_j: the distance between the transmitting node *j* and the next hop node.

BER_j: the estimated BER for the link between node *j* and the next hop node.

EDR_j: the expected data rate for the link between node *j* and the next hop node.

$CEC_{j,s}$: the current energy consumption between node *j* and node *s*, before the topology change.

$DR_{l,s}$: the current data rate that node *j* has through link *l* to node *s*, before the topology change.

$P(n)$: set of nodes that are members of the solution space and can collaborate with node *j* for the topology change (e.g., x, y in Figure 5-4).

e: number of UEs that participate in the multi-hop relay.

Phases 1 and 3 are applied by the head of the cluster, while UE_0 and the rest UEs of the search space undertake to apply phase 2. Below we present the mechanisms for the calculation of the H_{stop} (i.e., phase 1) and the recursive algorithm that is used to build local solutions spaces (i.e., phase 2).

The formation of the solutions space is terminated either when there is no UE that could reduce the *EEC* below the *CEC* ($EEC_{e2e} \leq CEC$) or when the number of hops from the initiating node UE_0 has reached its maximum value, denoted by H_{stop}. The head of the cluster calculates the stopping criterion (i.e., the value of H_{stop}), taking into account its local topology. Specifically, it considers: a) OF_r, which is the *OF* of the sub-graph that is formed by AP *r*, where UE_0 is associated with and the neighboring APs of AP *r*, b) the number of UEs that exist in the context of the cluster, denoted by *m*, and c) the total number of UEs that are associated with the APs that are neighbors of AP *r*, denoted by *z*. We calculate H_{stop} as follows:

$$H_{\text{stop}} = \left| h_{\max} \left(1 - \frac{z}{m}\right)^{OF_r} \right| \tag{5-25}$$

where h_{max} is the threshold for the maximum number of hops (e.g., h_{max} =5). H_{stop} is inverse-proportional to the number of UEs that are close to UE_0. A large number of neighboring UEs indicates that there are more networking opportunities, without increasing the size of the path and consequently H_{stop}. Low OF_r shows that the APs are sparsely placed and consequently their distance is large. Hence, the further extension of the search space, in terms of hops, does not provide more opportunities for an energy efficient multi-hop path, since the goal is to create a (multi-hop) path that satisfy: $EEC_{e2e} < CEC$.

After the specification of the search space, the collaborative procedure for the creation of the local solutions space of each UE_j, denoted by S_j, takes place. Let $S_j = \{S_{j,1}, S_{j,2}, S_{j,3}, ..., S_{j,\zeta}\}$, $\zeta \in \aleph$ denote ζ (multi-hop) paths of UE_j solutions space. The root of each $S_{j\zeta}$ is UE_0 and it terminates at an AP, while all intermediate nodes along a multi-hop path are UEs. At this point we should clarify that the development of the solutions space starts from UE_0, which then identifies the neighboring UE nodes that can extend the solutions space. This process is repeated until the H_{stop} criterion is satisfied. Let $P_j = \{P_{j,1}, P_{j,2}, P_{j,3}, ..., P_{j,\xi}\}$, $\xi \in \aleph$ denote the set of UEs that are sensed by UE_j ($P_j \subseteq U_{UE_j}$) and undertake to extend the solutions space after UE_j. The scheme that each UE_j uses in order to build S_j and P_j is described in Table 5-3.

Table 5-3: Algorithm for UE building of local solutions space

```
UE node: BuildMultiHopPath ($UE_j, L_j, EEC_j$)

Build $A_{UE_j}$ and $U_{UE_j}$ of $UE_j$, discovering its one hop away nodes
For each $AP_i \in A_{UE_j}$
    Calculate $EEC_j^{AP_i}$, $EDR_j^{AP_i}$
    $EEC_{e2e}^{AP_i} = EEC_j + EEC_j^{AP_i}$
    If $EEC_{e2e}^{AP_i} < CEC$ and $EDR_j^{AP_i} < CDR_{j,s}$
        $S_j \leftarrow \{ L_j, AP_i, EEC_{e2e}^{AP_i} \}$
    End
End
Send $S_j$ to the Cluster Head;
If length($L_j$) < $h_{\text{stop}}$ then
    For each $UE_g \in U_{UE_j}$
        Calculate $EEC_{j,g}$, $EDR_j^{UE_g}$
        $EEC_{e2e}^{AP_i} = EEC_j + EEC_{j,g} + EEC_g^{AP_i}$
        If $EEC_{e2e}^{AP_i} < CEC$ and $EDR_j^{UE_g} < CDR_{j,s}$
            $P_j \leftarrow \{ L_j, UE_g, EEC_{e2e}^{AP_i} \}$
        End
    End
    For each $UE_k \in P_j$
        Call BuildMultiHopPath ($UE_k, \{L_j, UE_k\}, EEC_j + EEC_{j.k}$);
    End
End
```

UE₀ is the first node that applies the algorithm of Table 5-3, and the initial input parameters are set as follows: $S_j = \varnothing$, $EEC_{e2e} = 0$ and $L_j = UE_0$, where L_j denotes a temporary structure of UE_j, which keeps the current (multi-hop) path that has been built, starting from the root (i.e., UE_0). UE_j discovers one-hop away APs and UEs building A_{UE_j} and U_{UE_j}, respectively. The response messages that UE_j receives from A_{UE_j} and U_{UE_j} nodes includes the estimated distance between UE_j and $UE \in U_{UE_j}$, the monitored BER, as well as the estimated distance between UE_j and $AP \in A_{UE_j}$. All this information is necessary for the calculation of UL *EEC* and *EDR* for this hop. Furthermore, each

$UE \in U_{UE_j}$ indicates its availability to participate in the multi-hop formation, based on its battery level. After the discovery phase, UE_j calculates for each sensed AP (AP_i) the $EEC_{e2e}^{AP_i}$, using (5-23) and the $EDR_j^{AP_i}$. If both $EEC_{e2e}^{AP_i}$ and $EDR_j^{AP_i}$ are below *CEC* and *CDR* respectively, then the path from UE_j to AP_i is included in the solutions space of UE_j (i.e., S_j). The solutions space of each UE_j is transmitted to the head of the cluster. Thereinafter, having checked all sensed APs, UE_j calculates *EDR* and EEC_{e2e} between UE_j and all available neighboring UEs ($UE_g \in U_{UE_j}$). In this case, for EEC_{e2e} estimation we consider as the ending point the AP where UE_g is associated with (i.e., AP_i). Similarly, if both $EEC_{e2e}^{AP_i}$ and *EDR* is below *CEC* and *CDR*, correspondingly, then the path from UE_j to UE_g is included in P_j set. Then, each UE of P_j undertakes to repeat the above procedure, provided that the distance from UE_0 is not more than H_{stop} hops. The goal is to extend the existing solutions space, by including each $UE_k \in P_j$ that leads to an AP that reduces the UL energy consumption of UE_0.

The total solutions space of UE0 that the head AP builds, using the transmitted local solutions spaces of UEs, includes all candidate paths. UE_0 is the root node of the total solutions space and APs are the leaf nodes. The total solutions space could be described as a directed acyclic graph (e.g., Figure 5-5), where (5-24) is applied in order to decide on the most appropriate path.

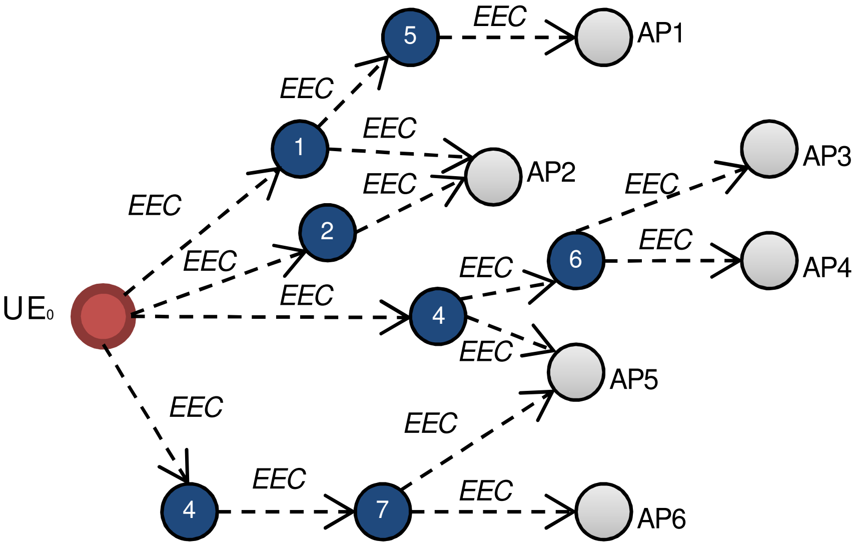


Figure 5-5: Sample solutions space that head AP builds for UE_0

Table 5-4: Algorithm for AP channel reselection

```
Cluster Head: Channel Reselection
1:  For each $AP_i, i = 1...n$
2:      Calculate $RC_{AP_i}$;
3:  End
4:  Select $AP_i$ with the maximum $RC_{AP_i}$ value;
5:  $A_{AP_i} \leftarrow$ Set of APs that are sensed by $AP_i$
6:  For each Candidate Channel $Ch_q$;
7:      For each neighboring $AP_k \in A_{AP_i}$
8:          $f_2 = Overlap(Ch_q, Ch_{AP_k})$;
9:          $f_3 =$ number of UEs that belong to $\bar{A}_{UE_j}$ and sense $AP_k$;
10:         $RC_{AP_i}^{Ch_q} = RC_{AP_i}^{Ch_q} + f_2 \cdot \left( f_3 \cdot T_{AP_k}^{Tx} + T_{AP_k}^{Tx} \right)$
11:     End
12: End
13: Select Channel $Ch_q$ for $AP_i$ that leads to the minimum $RC_{AP_i}^{Ch_q}$
14: $RC_{AP_i} = RC_{AP_i}^{Ch_q} - \left( uT_{AP_i}^{Tx} + \sum_{j=1}^{u} \left( f_{UE_j} T_{UE_j}^{Tx} \right) \right)$
15: Select $AP_i$ with the next maximum $RC_{AP_i}$ and GoTo Step 5
```

5.4.3 Channel Reselection

The head of the cluster triggers the channel reallocation phase in order to reduce the consumed energy for the reception and sensing of data packets due to the overlap of the selected channels among neighboring APs. The head identifies those APs (i.e., cells) that receive a high number of packets from adjacent APs and then checks whether there is a more appropriate channel to select that will reduce the consumed energy for the reception and processing of “redundant” packets. The channel selection process per AP takes into account the transmitted traffic from neighboring APs (i.e., cells) and the overlap between the available channels. For the identification of the APs that receive a high volume of data from neighboring cells the Reception Cost function (RC) is introduced and calculated by the head for each AP of the cluster for each member AP_i. For RC_{AP_i} we consider the total data traffic (in b/s) that is received by AP_i ($T_{AP_i}^{Rx}$) and the traffic received by all UEs ($UE \in \bar{U}_{AP_i}$) that are associated to AP_i ($T_{UE_j}^{Rx}$). In RC_{AP_i} we do not consider the data traffic that is transmitted from both AP_i ($T_{AP_i}^{Tx}$) and UEs that are in the same cell (i.e., associated with AP_i), because this type of traffic cannot be avoided regardless of the channel that AP_i has selected. Hence RC for AP_i is calculated as follows:

$$RC_{AP_i} = \left(\sum_{j=1}^{u} T_{UE_j}^{Rx} + T_{AP_i}^{Rx} \right) - \left(uT_{AP_i}^{Tx} + \sum_{j=1}^{u} \left(f_{UE_j} T_{UE_j}^{Tx} \right) \right) \tag{5-26}$$

where f_{UE_j} denotes the number of UEs are in the $\bar{U}_{UE_j}$ set and u the number of UEs that are associated to AP_i. The head of the cluster calculates RC_{AP_i} for all member APs and then according to the scheme that is presented in Table 5-4 concludes to the channel (Ch_q) for each AP_i that will minimizes its RC function and consequently the consumed energy for the reception of packets that arrive from adjacent cells. For the reduction of RC the existing topology graph, the traffic of neighboring and the overlap among channels are taken into account. $Overlap(Ch_q, Ch_{AP_k})$ provides the level of overlap between channel Ch_q and the channel that AP_k uses. An example of this function is provided in [142].

5.5 Performance Evaluation

The mechanisms presented in sections 5.3 and 5.4 for energy efficiency and CCO as well as the cluster-based self-organization scheme have been implemented using the OPNET Modeler v14.0 simulation software [143]. Performance results for energy consumption in the communication component at the AP side and the UE side are described and discussed. Both the AP node and the user node support the IEEE 802.11b/g wireless LAN station model (i.e., WLAN Receiver/Transmitter, WLAN MAC layer). We have implemented the SON modules for the cluster formation, the energy saving and the CCO mechanisms presented in sections 5.2, 5.3, and 5.4. UEs are monitoring points and enforce the decided re-configuration actions, e.g., handover, channel re-selection. During the simulation, all UEs are fixed and their handover capability is enabled.

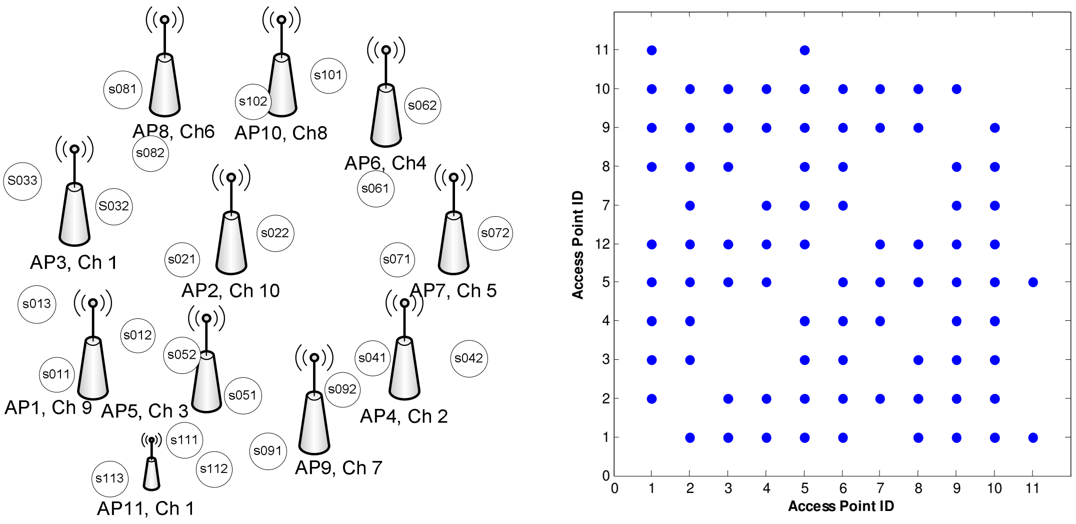
All APs have the same transmission power (0.00035 W), apart from AP 11 (Figure 5-6-a) whose transmission power is set to 0.00005 W in order to create a smaller cell area (~50 meters), for reasons explained below. UE transmission power is adapted based on the distance from the associated AP. For the exchange of monitoring data among the member APs and the head AP in the context of the cluster, a second physical radio

Table 5-5: Simulation parameters

Parameters	Value
PHY	IEEE 802.11 b/g
Frequency (MHz)	2.401
Bandwidth (kHz)	22
AP Transmission Power (W)	0.00035
Packet Reception-Power Threshold (W)	7.33 E-14
Maximum Transmission Range	220m
Physical Characteristics	Direct Sequence
Maximum Data Rate (bps)	11.000.000
Modulation	bpsk
CTS/RTS	Disabled
Packet Payload size (bits)	1152
Efs (nJoules/bit)	50.0
Elec (pJoules/bit/m^2)	10.0
w_1	0.30
w_2	0.70
B_r	0.70
B_d	0.25
B_e (nJ/bit)	0.50

channel has been established per node, adopting the Cognitive Control Radio (CCR) concept [144]. This channel does not affect the data traffic that UEs and APs exchange. The values of the simulation parameters are provided in Table 5-5.

We have chosen a network topology that consists of eleven APs and twenty four UEs. The initial established network topology is presented in Figure 5-6-a, while Figure 5-6-b shows the adjacency matrix of the APs. A dot in the adjacency matrix denotes that AP IDs correspond to APs that are in the transmission range of each other. Moreover, Figure 5-6-(a) illustrates the unique ID of APs and UEs, as well as the channel that has been allocated with. The first two digits of the UE ID indicate the ID of the AP the corresponding UE is associated with. AP2 is the head AP and periodically collects measurements from all member APs in the cluster. Each UE provides monitoring reports to the AP it is associated with. In the simulation, we inject both UL traffic from the UEs towards the AP and DL broadcast traffic from the APs towards the associated UEs. The UL traffic level is higher than the DL traffic. Transmitted data packets (Tx) are also received (i.e., sensed) by those neighboring nodes that use overlapped frequency channels, apart from the destination nodes that exist in the same cell. At this point we should note that the term received traffic (Rx) denotes the data traffic that is sensed by



(a) (b)

Figure 5-6: Simulation topology (a) Network graph (b) APs adjacency matrix

each node (UE, AP) regardless if it is the destination node or not. The chosen network topology as well as the configuration of the traffic generators and the network nodes (e.g., AP 11 transmission power) allows us to evaluate in the same simulation scenario all states and conditions for energy saving that have been discussed in section 5.3. We have configured the duration of each simulation run to 7 minutes.

Initially, we present the energy consumption of UEs and APs when the energy saving solution is not enabled. The purpose of this simulation is to present the participation of each phase (Rx, Tx) and node type (UE, AP) in the energy consumption of the cluster area. The total traffic that APs and UEs transmit and receive per second, for the duration of the simulation, is illustrated in Figure 5-7. Figure 5-8-a and Figure 5-8-b present the energy consumption (in Joules/sec) for alls APs and UEs in the cluster, for the case that the proposed CCO solution is not enabled. For the energy consumption calculation we are focusing on the communication component, and we have used the formulas presented in section 5.2.

For the duration of the simulation, 93% of the total energy (5.1 kJ), in the communication component of the APs, is consumed during the transmission phase, while only 0.39 kJ is required for the reception phase (Figure 5-9). On the other hand, UEs consume 0.98 kJ (69%) for the reception of the data packets, and 0.43 kJ (31%) for the transmission phase. This is (partly) due to the shorter total transmission range of the UEs compared to that of the APs. A small part of the consumed energy for data from the MAC layer control packets that APs and UEs exchange (e.g., beacon packets). Both Rx and Tx phases affect significantly the energy consumption of UE, while for AP the Tx phase is the most energy consuming.

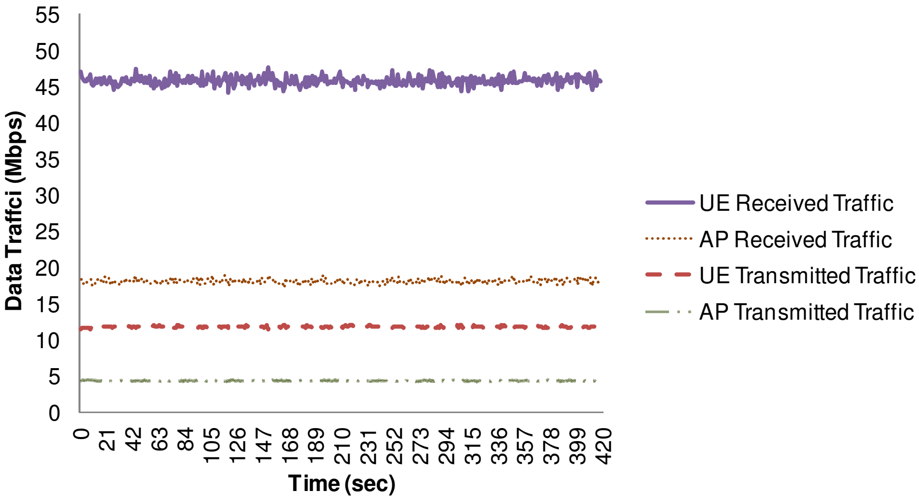
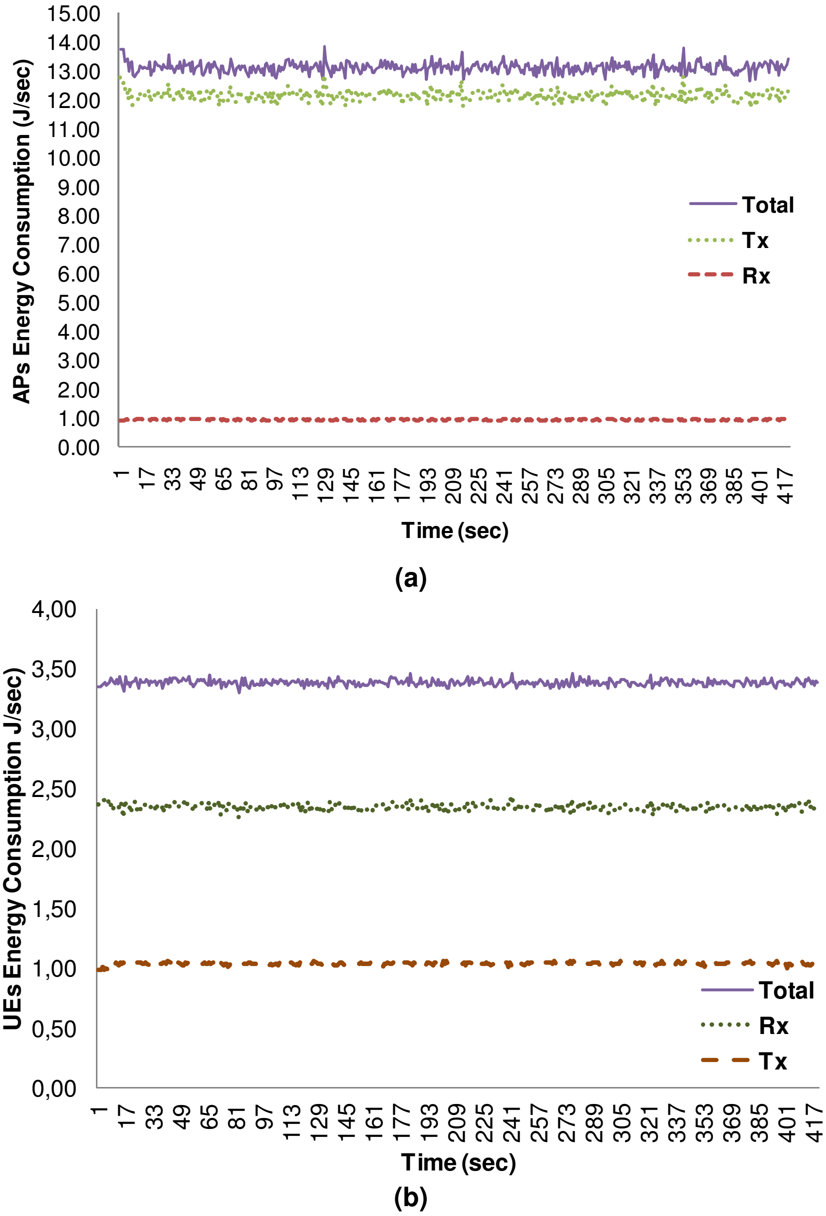


Figure 5-7: Cluster area APs and UEs data traffic – Disabled energy saving and CCO



(a)

(b)

Figure 5-8: Disabled energy saving and CCO: (a) APs energy consumption (b) UEs energy consumption

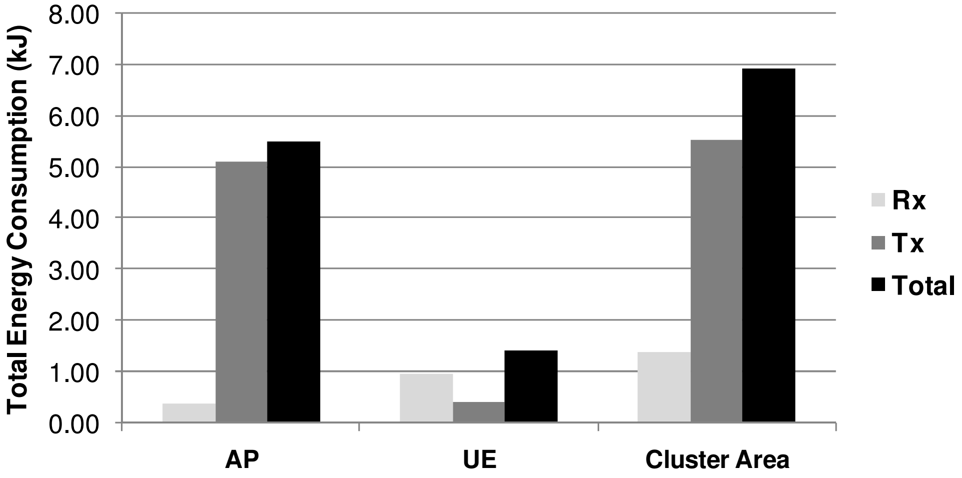
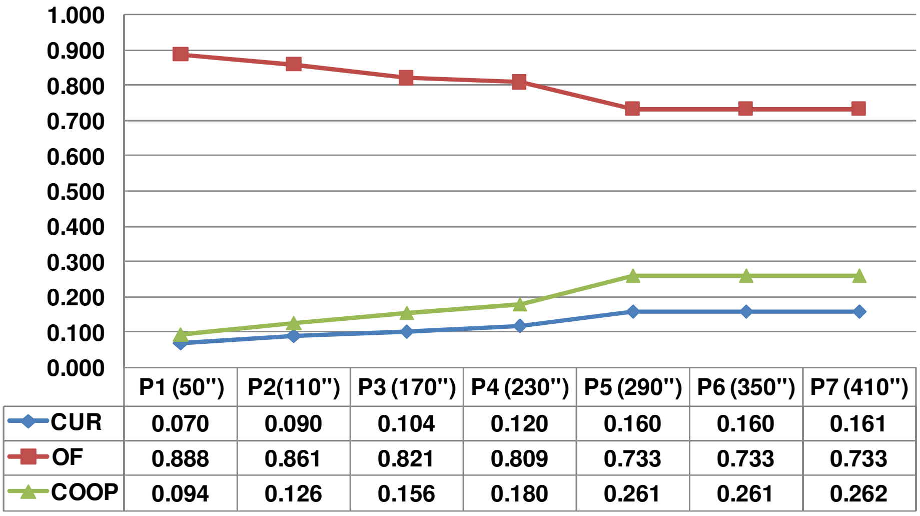


Figure 5-9: Total energy consumption – Disabled energy saving and CCO

Then we repeat the experiment, having enabled the proposed solutions described in section 5.4 for energy saving and CCO. The purpose of this simulation run is to evaluate the performance of the proposed SON algorithms and identify the change of energy consumption among the different type of nodes and components, after each configuration. The initial network topology and the configuration of the traffic generators is exactly the same as in the simulation run presented above. AP2 collects from the cluster members their topology graph as well as their status (load, received data, transmitted data etc). The head of the cluster, using the collected information from all member APs, calculates periodically (i.e., every 60 seconds) the COOP metric and the changes in energy consumption ($\Delta EC_{AP}^{Rx}, \Delta EC_{UE}^{Rx}$, ΔEC_{AP}^{Tx}, ΔEC_{UE}^{Tx}). Then the proposed rule-based decision making scheme (Table 5-1), checks whether the selection of an adaptation action is necessary for energy saving or resource usage optimization.



	P1 (50")	P2(110")	P3 (170")	P4 (230")	P5 (290")	P6 (350")	P7 (410")
CUR	0.070	0.090	0.104	0.120	0.160	0.160	0.161
OF	0.888	0.861	0.821	0.809	0.733	0.733	0.733
COOP	0.094	0.126	0.156	0.180	0.261	0.261	0.262

Figure 5-10: Cluster-level CUR, OF and COOP values during APs deactivation phase

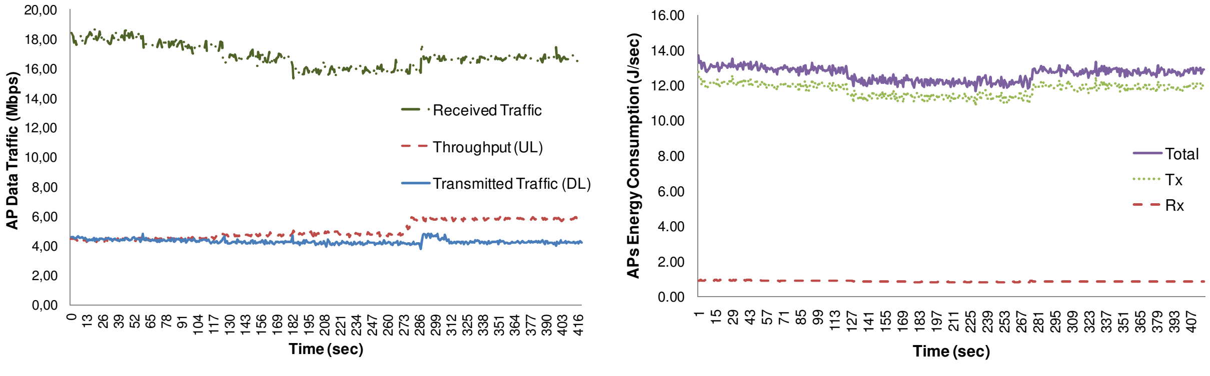
Table 5-6: APs local COOP

Period	AP1	AP2	AP3	AP4	AP5	AP6	AP7	AP8	AP9	AP10	AP11
P1 (50")	0.130	0.123	0.232	0.032	0.016	0.116	0.077	0.032	**0.007**	0.031	0.032
P2 (110")	0.148	0.152	0.286	0.056	**0.021**	0.132	0.091	0.055	-	0.043	0.048
P3 (170")	0.196	0.122	0.327	0.062	-	0.167	0.101	0.052	-	0**.035**	0.062
P4 (230")	0.183	0.135	0.308	0.064	-	0.173	0.109	**0.047**	-	-	0.064

Table 5-7: APs deactivation and UEs handover

Period	AP deactivation	UEs Handover	
		UE ID	*new AP ID*
P1 (50")	AP 9	S091	AP 5 (ch3)
		S092	AP 4 (ch2)
P2 (110")	AP 5	S051, s052	AP 1 (ch9)
		S091	AP 4 (ch2)
P3 (170")	AP 10	s101	AP 6 (ch4)
		s102	AP 8 (ch6)
P4 (230")	AP 8	S081, s082	AP 3 (ch1)
		s102	AP 2 (ch10)

Figure 5-10 presents the values of CUR, OF and COOP that AP2 calculates periodically. In addition DL and UL are used for the calculation of the CUR. We notice that for the first four periods the calculated COOP value is below the set threshold (*B_d=0.25*). Hence, the scheme for APs deactivation, described in Table 5-2, is triggered. The network topology and the received or sent data traffic during the simulation run, changes as a result of APs deactivation. AP that has the minimum local COOP value is selected for deactivation (Table 5-7), while Table 5-6 presents the AP that is selected for deactivation in each period and the APs where the associated UEs are handed over.

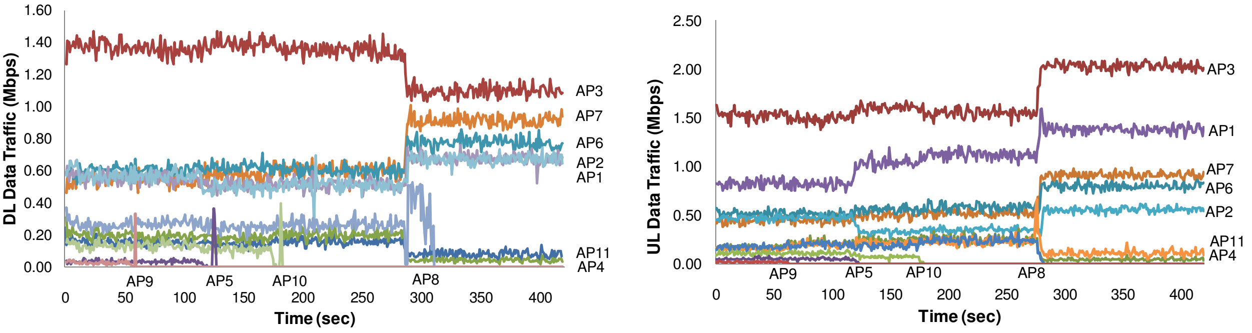


(a) (b)

Figure 5-11: APs deactivation enabled (a) Cluster area APs total data traffic (b) APs energy consumption

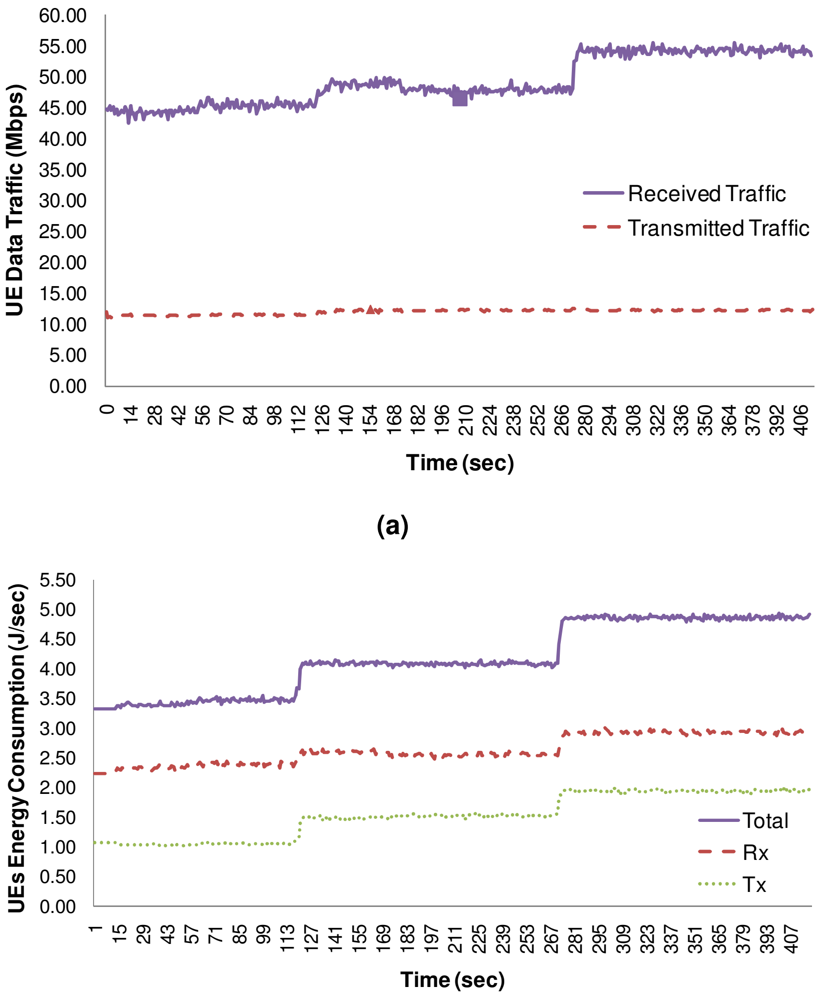
Figure 5-11-a shows the variation of the cluster-level traffic of APs, while Figure 5-12-a and Figure 5-12-b show the DL and the UL traffic of each AP during this simulation run. We notice that AP 9, AP5, AP10, and AP8 reception and transmission capability is set to zero, after their switch-off, at the respective time instance (i.e., T 55, T 122, T 180 and T 270). The modification of the network topology and configuration increases the throughput of APs (i.e., UL traffic) from 4.42 Mbps to 5.82 Mbps, while there is a slight reduction in AP DL traffic, from 4.45 Mbps to 4.27 Mbps. However, the deactivation of the RF interface of the four APs leads to a reduction of the totally sensed traffic for the APs side from 18.14 Mbps to 16.71 Mbps. Hence, less energy is consumed totally in the cluster for AP reception of data packets; EC_{AP}^{Rx} is reduced from 0.94 J/sec to 0.86 J/sec (Figure 5-11-b). According to (18), ΔEC_{AP}^{Rx} in the cluster is reduced from 57.46 nJ/bit to 50.00 nJ/bit (Figure 5-14-a). We have a 13% decrease of the consumed energy for APs reception of data traffic that is transmitted by neighboring UEs ($T_{UE_j}^{Tx}$) and APs ($T_{AP_i}^{Tx}$). We also note that further energy saving is achieved if we consider the reduction of the consumed energy from other modules and components that are de-activated at the AP side.

The slight reduction that appears in the EC_{AP}^{Tx}, from 12.12 J/sec to 11.93 J/sec, after AP de-activation (Figure 5-11-b), is due to APs DL data rate reduction, mentioned above (Figure 5-11-a). After UEs handover, there is no change in the transmission distance of the APs, and, as expected, the consumed energy for each bit that the APs transmit remains the same (2428 nJ/bit).



(a) (b)

Figure 5-12: APs deactivation enabled (a) APs DL data traffic, (b) APs UL data traffic



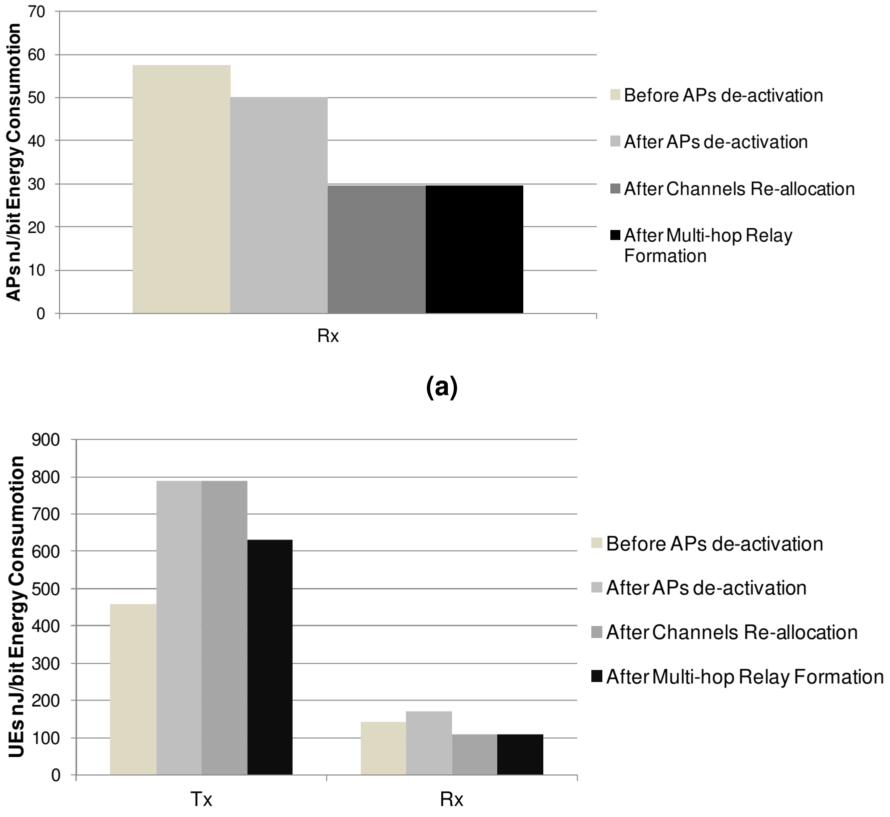
(a)

(b)

Figure 5-13: APs deactivation enabled (a) Cluster area UEs data traffic (b) UEs energy consumption

UE energy consumption for both data packet transmission and reception (EC_{UE}^{Tx} , EC_{UE}^{Rx}) has been significantly increased (Figure 5-13-b), after the finalization of APs deactivation and UEs handover process (Time instance: T 280). This is because after their handover the distance of the UEs from the associated APs is in many cases longer; this leads to the increase of the power required for the transmission of data packets. Furthermore, the concentration of more UEs in close frequency channels (in the post-handover phase) increases the sensed traffic that derives from adjacent cells (Figure 5-13-a). The latter affects negatively the EC_{UE}^{Rx} metric. The head AP, according to the proposed rule-based decision making scheme (Table 5-1), checks ΔEC_{UE} and ΔEC_{AP}. The obtained results show that after the finalization of APs deactivation phase ($COOP = 0.28$), we have $\Delta EC_{UE}^{RX} = 26.98$ nJ/bit (Figure 5-14-b). Since $\Delta EC_{UE}^{RX} \geq B_e$ (Table 5-1), the head AP triggers the channel reallocation adaptation action in order to reduce UE energy consumption from the reception of data traffic that comes from neighboring cells.

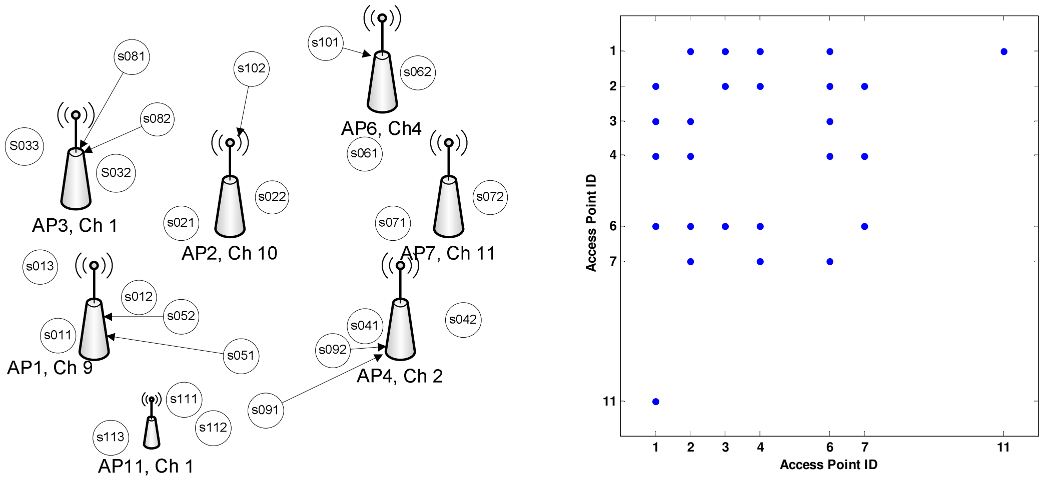
For the re-allocation of the channel that each AP uses, the head AP uses the scheme described in Table 5-4. The data traffic that each node receives from adjacent cells, as well as the traffic that comes from the nodes (UE, AP) of the same cell, are collected by the head AP in order to calculate the RC of each AP in the cluster, using (5-26). The proposed scheme takes into account the topology graph, the overlap of the selected channels (Figure 5-15), and the traffic levels after the APs de-activation phase (see T 290 in, Figure 5-10). At this point, we should note that the f_2 parameter is set to 1 if there



(a)

(b)

Figure 5-14: Energy consumption in nJoules/bit (a) APs Rx phase, (b) UEs Tx and Rx phases



(a) **(b)**

Figure 5-15: (a) Network graph after APs de-activation (b) Updated APs adjacency matrix

is even the minimum possible level of overlap between two channels (e.g., Channel 1, Channel 5). Otherwise, if there is no overlap among two frequency channels (e.g., Channel 2, Channel 7) then the f_2 parameter is set to 0.

The rank of APs in ascending order, according to the calculated RC value, is the following: AP2, AP7, AP6, AP1, AP 4, AP 11 and AP3. AP2 and AP7 have the highest RC value because they sense the largest volume of data packets from neighboring cells. The head of the cluster checks whether the change of the selected channel per AP could reduce its RC cost. After applying the proposed scheme, only AP7 selects a different channel: AP 2: Channel 10, AP 7: Channel 11, AP 6: Channel 4, AP 1: Channel 9, AP 4: Channel 2, AP11: Channel 1, AP 3: Channel 1.

Then we repeat the above experimentation scenario using the formed topology of Figure 5-15 and enforcing the decided channel re-allocation action. This results to the reduction of the data traffic that both APs and UEs receive from adjacent cells, totally in the cluster area (Figure 5-16). Consequently less energy is consumed for the reception phase. UEs and APs energy consumption measured in nJs/bit is improved by 35.7% and 40.5%, correspondingly (Figure 5-14-a, Figure 5-14-b). As expected the consumed energy for the Tx phase is not affected.

After the APs deactivation period, as illustrated in Figure 5-17-a, and still after the reallocation of frequency channels, the average consumed energy for UEs transmission phase has been substantially increased ($\Delta EC_{UE}^{Tx} = 331.232$, Figure 5-14-b). Almost all UEs after the handover phase are associated with APs that are at a longer distance from the initial AP (e.g., S91, S52, and S82). In this case, the rule-based decision making scheme

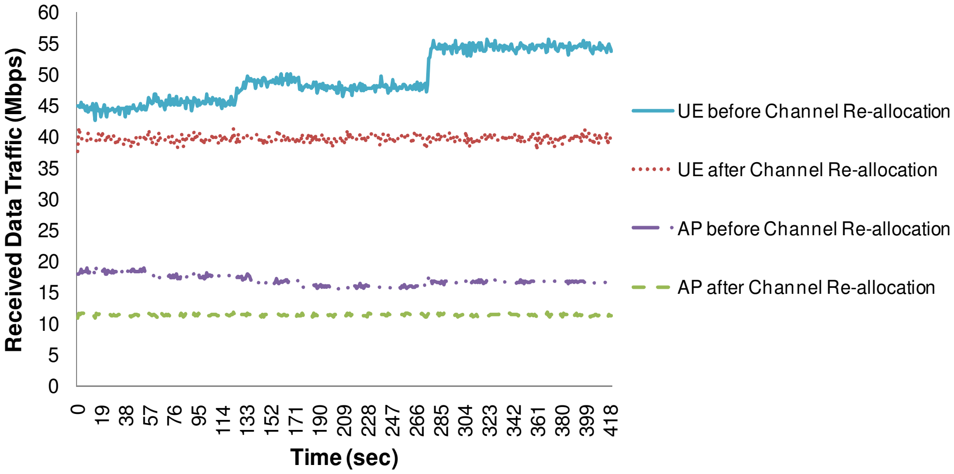
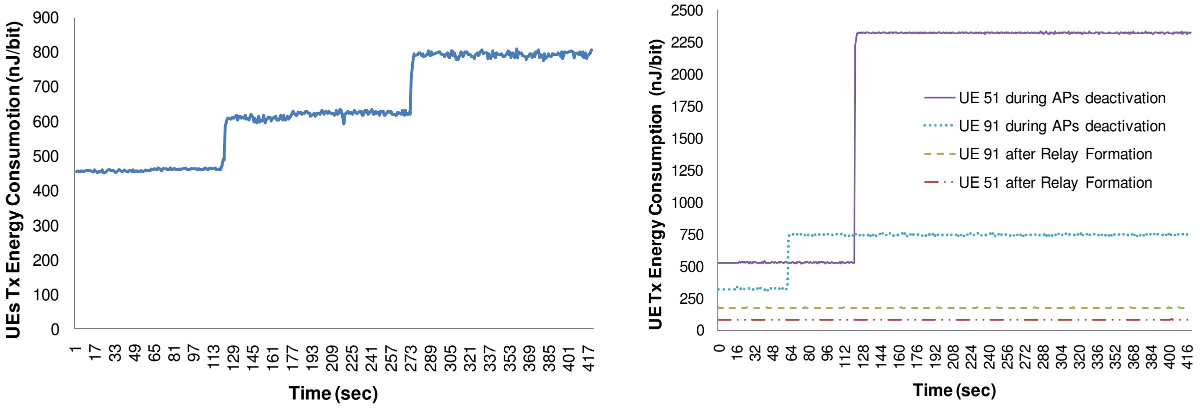


Figure 5-16: All UEs and APs received data traffic before and after channel re-allocation



(a) (b)

Figure 5-17: Tx Energy consumption (a) All UEs during APs de-activation simulation (average) (b) UE S51 and S91 before and after relay formation

of the head AP (Table 5-1) leads to the formation of a shorter (multi-hop) link for each UE that causes an increase of the EC_{UE}^{Tx} and has high levels of UL traffic.

The UEs that show a substantial increase of their energy consumption for the transmission of data packets ($\Delta EC_{UE}^{Tx} > B_e$), which the head AP has selected, are the following: S51, S52, S91, S92, and S82. AP2, based on the algorithm presented in section 5.4.3, investigates the building of a multi-hop path towards an AP, calculating firstly the search space from the initiating UE per case (H_{stop}). Considering the existing topology and the neighboring nodes (UEs, APs) that each UE can sense, only UEs S91 and S51 manage to form a multi-hop path that reduces their consumed energy for the transmission phase.

Applying equation (25) and taking into consideration the topology of Figure 5-15, the H_{stop} value is one hop for both S91 and S51. Table 5-8 gives the values of the parameters used for the calculation of the H_{stop} for each UE. The energy consumption of UEs S91 and S51, in nJoules/bit, before the multi-hop relay formation and during APs de-activation phase is presented in Figure 5-17-b. We notice that after the deactivation of AP 9, at time instance T55, the EC_{UE}^{Tx} of S91 increases, from 322.67 nJoules/bit to 749.38 nJoules/bit, due to the handover of S91 to AP 4. In addition, the deactivation of

Table 5-8: Estimation of S91 and S51 search space (H_{stop})

UE ID	AP *r*	h_{max}	m	z	OF_r	H_{stop}	$\lvert H_{stop} \rvert$
S51	AP 4	4	24	17	0.9	1.31	1
S91	AP 1	4	24	19	0.8	1.14	1

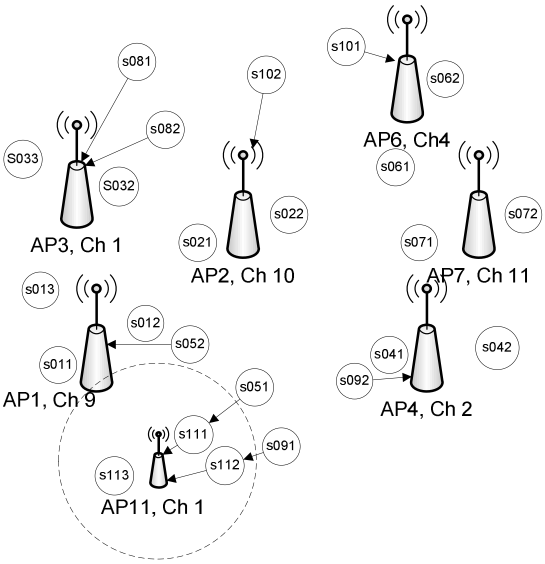
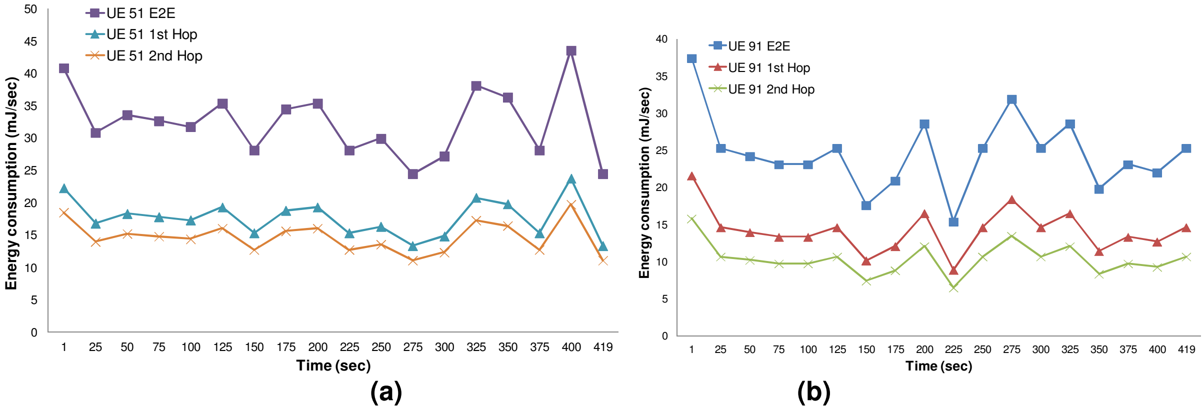


Figure 5-18: Network graph after S91 and S51 two-hops relay formation

AP 5 at time instance T119 leads to the increase of EC of S51 from 530.48 nJoules/bit to 2325.19 nJoules/bit. Thus, the CEC of UEs S91 and S51, is 749.38 nJoules/bit and 2325.19 nJoules/bit, respectively. The CDR is set to 0.50Mbps. The application of the scheme presented in section 5.4.3 leads to the formation of the two hops link for S91 (S91-S112-AP11) and S51 (S51-S111-AP11), since all other neighbors of both UEs are in a longer distance and, therefore, energy improvement is not achieved. The topology after the establishment of the multi-hop paths is presented in Figure 5-18.

Figure 5-17 shows the reduction of S91 and S51 energy consumption for the transmission phase, in nJoules/bit. Having formed the relay, the EC_{UE}^{Tx} of S91 and S51 is reduced due to the formation of the 2 hops path and the shorter end-to-end distance. On the other hand, S112 and S111 energy consumption is increased due to the traffic they undertake to forward. However, the cluster level energy consumption for EC_{UE}^{Tx} is reduced (Figure 5-14-b). This is because the end-to-end (i.e., two hops) consumed energy



(a) (b)

Figure 5-19: E2E UL energy consumption for (a) UE S51 and (b) S91

starting from S51 and S91 is lower than the energy that both UEs consumed for the transmission, before the formation of the relays. For the calculation of the end-to-end energy consumption, we consider the EC_{UE}^{Tx} of the first hop (S51, S91) and the second hop nodes (S112, S111). Figure 5-19 presents the end- to-end consumed energy as well as the consumed energy for each hop of the relays. The average data traffic that S91 and S51 transmit is 80.5 Kbps and 126.5 Kbps, respectively.

The energy that the neighboring nodes (UE, AP) consume for the reception (i.e., sensing) of the data packets (EC_{UE}^{Rx}, EC_{AP}^{Rx}), which are sent from the nodes of the relay, is not higher comparing to the energy that was required in the previous topology (i.e., before the formation of relays). The transmission distance from the initiating nodes (S51 and S91) to reach an AP is smaller in both cases; hence fewer nodes are inside the transmission radius to sense the transmitted data packets from S51 and S91. We should note that the creation of the two hops relay has an impact on the average end-to-end delay, since UEs S112 and S111 process the packets that they receive and forward. Moreover, the data rate of the second hop, in both cases is lower from the Tx rate of S91 and S51 leading to the reduction of the total data rate and to the dropping of data packets.

At this point some of the key outcomes of the analysis should be noted:

- The adoption of a cluster-based approach for energy management provides awareness for the interactions and the dependencies among various nodes, improving the effectiveness of energy saving solutions.
- The AP RF deactivation reduces the energy that is consumed in the AP reception phase. However, it increases the energy consumption of UEs for the reception and the transmission phase under specific topology and traffic conditions (e.g., high overlap of frequency channels, UL/DL ratio).
- Channel reallocation facilitates energy saving at both the AP and UE for the reception phase, eliminating the traffic that comes from neighboring cells.
- The formation of UEs multi-hop relays improves the energy consumption in the data packet transmission phase, especially for the cases of high UL traffic levels.
- The coordination among the various configuration actions and the introduction of metrics that capture the energy consumption change in the different types of nodes (AP, UE) and component (Rx, Tx) are required in order to manage the complexity of energy saving in a wireless communication system.

CHAPTER 6

SELF-ORGANIZING NETWORKS COORDINATION: RESOLVING CONFLICTS AND DEPENDENCIES

Addressing the various CNMs (NECMs or NDCMs) in a SON as cognitive cycles that operate individually might lead to conflicts or ripple effects; thus affecting the stability of the network or even leading to the degradation of network performance. The performance metrics (PM) of the optimization or configuration problems of a SON and the associated configuration actions (CA) appear many interactions and dependencies.

The avoidance of conflicts either on the CAs (i.e., adaptations, re-configurations) or on the PMs and the joint assessment of the various SON problems is a challenging issue for SONs. A conflict in a control parameter (e.g., AP switch On/Off, antenna tilt) appears if two or more metrics, which exceed a predefined threshold, attempt to change the same CA towards different (i.e., conflicting) directions. On other hand, a conflict on a PM (e.g., interference, load) indicates that a triggered configuration action has a negative influence on another PM. This might lead to a sequence of adaptations creating a ripple effect situation. The avoidance of a conflict either on a parameter (i.e., adaptation) or on a metric and the joint assessment of the various SON problems is a challenging issue for SONs and specifically for a CNM. Conflicts on a CA or on a PM might appear in the context of the same device or among devices that exist in the same neighborhood.

In this chapter, we propose a novel scheme for the coordination of the various functions of a SON. We identify and resolve conflicts on PMs and CAs, which arise from the deduction phase of a CNM. The proposed scheme consists of three steps. Firstly, a time series-based approach is used in order to avoid checking a CA that is triggered by a PM which value appears continuous variations due to a temporary change of the network area. This helps the CNM to act proactively regarding conflicts and dependencies, avoiding the trigger of a possibly uncertain adaptation. In the next step, we check for conflicts on the (certainly) triggered CA. Only one direction per configuration action is prioritized, according to the urgency and the priority of the PM that triggers the corresponding CA. Finally, the impact of non-conflicting CAs on other PMs is analyzed based on performance metrics priorities and using a cost-benefit analysis approach. The goal is to select the CA with the highest priority that creates the fewer conflicts on PMs and has the minimum possibilities for the deterioration of other PMs. The proposed solution has been implemented and applied in a wireless network where

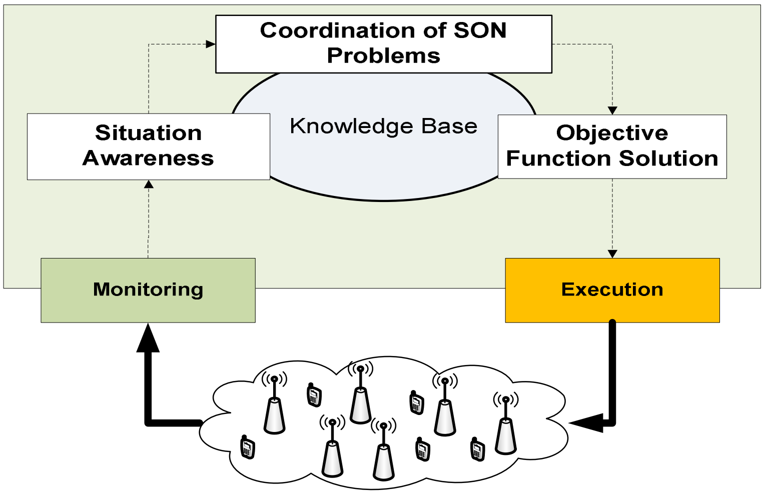


Figure 6-1: CNM and SON Coordination overview

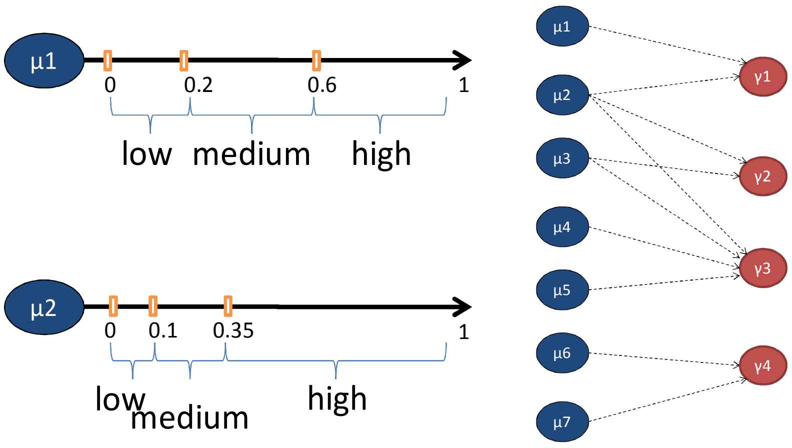
CCO, Interference and energy saving use cases are addressed by the SON Coordination scheme. Performance improvement has been evaluated deploying a wireless environment based on the OPNET simulation tool.

6.1 Network Model and Assumptions

Each CNM (NECM, NDCM) periodically retrieves measurements utilizing the monitoring tools of network elements and invokes the situation awareness phase. The latter based on the monitored values calculates all PMs and identifies performance improvement triggers (i.e., optimizations), according to the thresholds that the administrator has specified. For each trigger the situation awareness phase selects the appropriate CA, based on the introduced policy rules. If no action is required then the process is interrupted and the CNM waits for the next periodic invocation.

In the case that more than one CAs (i.e. adaptations) have been selected then the list of candidate CAs is transferred to the "Coordination of SON problems" functionality. In this phase the most appropriate CA is selected, according to the identified priorities as well as the impact of the triggered CAs and the related PMs. Finally, the objective function or the algorithm of the prioritized CA is solved and its outcome is executed by the mechanisms that reside in the respective network elements.

Let μ represent the PMs of the SON Coordinator with index $k \in \{1,2,...K\}$, where k is the total number of PMs. Moreover, γ denotes the available configuration actions with index $g \in \{1,2,...G\}$, where G is the total number of CAs. In the situation awareness phase, different monitoring parameters are jointly correlated to compose PMs that indicate how the CNM perceives its environment (e.g., load status, interference level). Based on the



(a) **(b)**

Figure 6-2: (a) Sample Performance Metrics numeric linguistic variables and sates, (b) Performance Metrics and Configuration Actions bipartite graph

importance of each PM for the operation of the network each PM μ_k belongs to a priority class $P(\mu_\kappa)$.

As it is described in section 3.3, the knowledge base is important for the operation of each CNM (NECM, NDCM), representing relationships among objects/events in order to support decision making tasks of the CNM. This knowledge base attempts to transfer the awareness and the knowledge that a human administrator has developed through experience or training. It keeps the policy rules and the thresholds denoted by $Th_{\mu_k}^{\gamma_g}$ for the triggering of a configuration action γ_g, assessing the values of PMs μ_k ($\mu_k \xrightarrow{Th_{\mu_k}^{\gamma_g}} \gamma_g$). Taking into account existing network condition, policy rules guide the network about which CA to use in order to move from one state to another. For instance, in the case of a high interference state a channel re-allocation action is triggered in order to reduce the interference levels and improve the performance of the network system.

However, the same CA might be triggered by more than one performance metrics (e.g., $\{\mu_1, \mu_2\} \rightarrow \gamma_1$ and $\{\mu_4, \mu_5\} \rightarrow \gamma_3$). Furthermore, the enforcement of a specific CA (e.g., $\mu_5 \rightarrow \gamma_3$) apart from the improvement of a specific PM, it might also affect other PMs that have not triggered its execution (e.g., CA γ_3 *affects* μ_2 PM). This fact complicates the decision making for the execution or not of a configuration and how the latter affects the whole system's behavior.

A bipartite graph that presents the CAs (C_i) that each M_i can trigger facilitates the identification of interactions and dependencies among CAs and PMs (Figure 6-2-a). This graph is used by the CNM and specifically by the SON coordination functionality in order to identify the direct and indirect relations among CAs e.g., the PMs that trigger a CA or the PMs that are affected by a specific CA.

6.2 Scheme for Conflicts and Dependencies Resolution

Below we present the scheme for the identification and the resolution of conflicts and dependencies among SON problems. This will lead to the selection of the appropriate CA for the case that more than one CAs have been triggered by the situation awareness phase. The three steps of the proposed algorithm are presented in Figure 6-3.

Define $C = \{C_1, C_2, \ldots C_n\}, n \in \Re$ as the set of the CAs ($C \subseteq \gamma$) that are triggered by the situation awareness phase of the NECM. $M = \{M_1, M_2, \ldots M_m\}, m \in \Re$ denotes the value of those PMs ($M \subseteq \mu$) that have triggered each $c_j \in C, j = 1, 2, \ldots n$. Sets *M* and *C* as well as the bipartite graph of PMs and CAs are the inputs to the algorithm for the analysis of SON problems interaction.

The SON coordination scheme that is presented below takes into account: a) the severity of each PM that triggers a CA as the latter is depicted in the priority list, b) the impact on other PMs, c) and the number of transitions (re-configurations) that take place in the communication system. Hence the goal of the SON coordination scheme is to solve the highest priority performance issue, with the least negative impact on other PMs via the lowest number of reconfigurations.

SON coordination requires the cooperation between the NECM and the NDCM as well as among NECMs of different devices (Figure 6-3). In the first step a scheme based on time-series is used for the detection of temporary fluctuations of the PMs that trigger a CA. Thus, we exclude from *C* those CAs that are triggered due to a temporal network condition and update *M* accordingly. Sets $C' \subseteq C$ and $M' \subseteq M$ are formed respectively. In the next step the candidate CAs (C') and the respective associated PMs (M') are checked in order to identify and resolve conflicts among the CAs of set C'. Excluding conflicting CAs, $M'' \subseteq M'$ and $C'' \subseteq C'$ sets for non-conflicting CAs are formed. The above steps take place both in the NECM and NDCM side. A CA that is selected by the NDCM is directed solved, since it has higher priority than the CAs of each NECM. This design decision is rational, taking into consideration that the NDCM monitors a greater network area, while the solution of a cluster-level problem will afterwards facilitate the handling of problems that have been locally detected. In the case that NDCM has not triggered any CA, the next step includes the coordination among neighboring NECMs

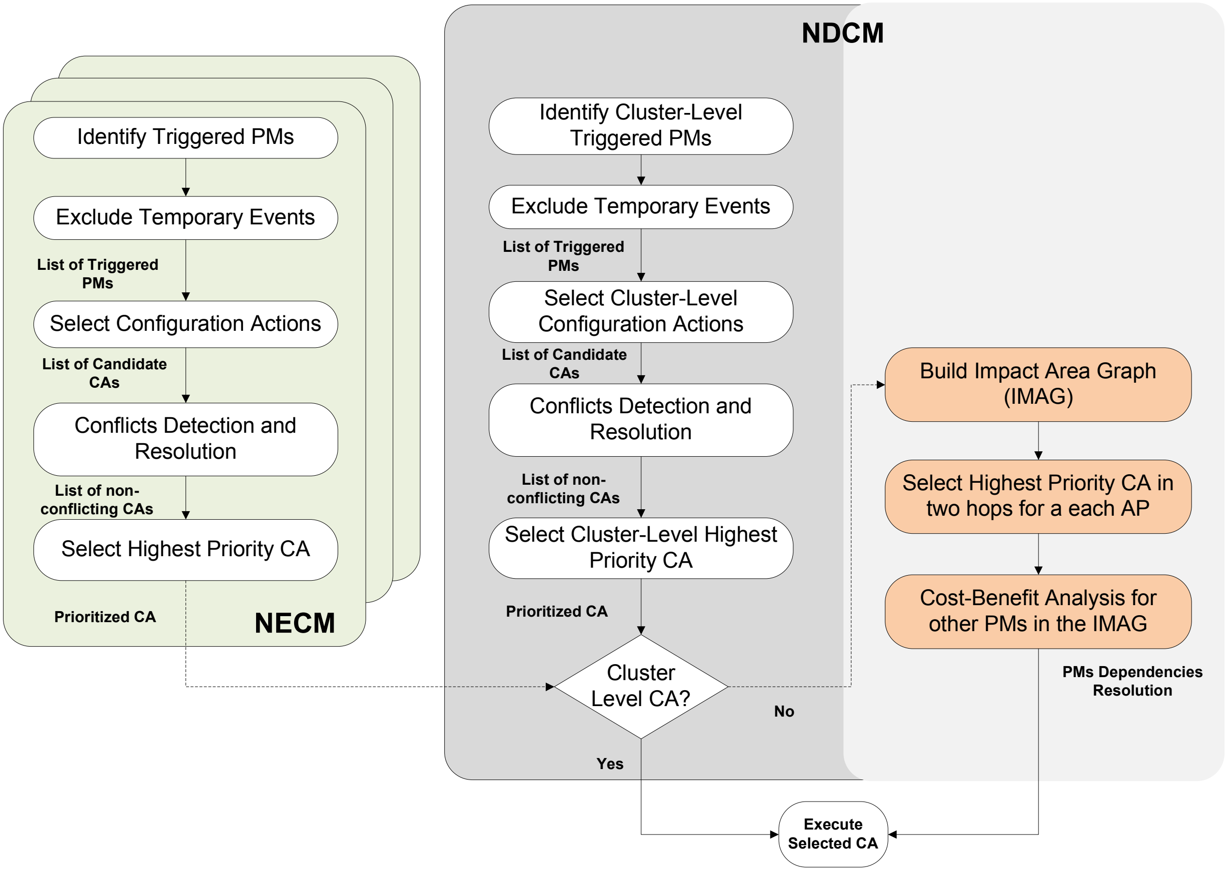


Figure 6-3: Scheme for the coordination of SON problems

(i.e. APs). Firstly, the impact area graph (IMAG) is built, which includes the highest priority CAs that each neighboring NECM intends to enforce. Then the highest priority action is identified in a two hops distance from each node. Thus, conflicts among neighboring nodes are avoided. Thereinafter, the prioritized CA is resolved using a Cost-Benefit analysis scheme in order to determine whether it is an appropriate decision, by estimating its impact on the other PMs.

The latter is crucial, since a significant deterioration of a PM might lead to the indirect triggering of further CAs (i.e., ripple effect phenomenon). After the end of this phase, the appropriate configuration action is prioritized and solved.

6.2.1 Temporary Effects of Performance Metrics

A short-term forecasting method has been used in order to exclude from list *C* those CAs that are triggered due to PMs' instant fluctuations or temporary values. This will protect the network from proceeding to needless adaptations that might affect its performance and stability. The SON coordination functionality considers $M_i \in M$ candidate for optimization in the case that the value of M_i continues to exceed the threshold that has triggered C_j ($M_i \xrightarrow{Th_{M_i}^{C_j}} C_j$) after *e* periodic cycles of operation.

We estimate the value of each $M_i \in M$ after *e* cycles, using a time series method that is based on double exponential smoothing method (Holt smoothing model). A smoothing method is designed to capture a trend, which estimates a smoothing equation when data has no seasonality or cyclicality. Having calculated M_i for the decision making time *t* ($M_{i,t}$), the M_i value for the next period is predicted by using the following equation:

$$M_{i,t+1} = a\left(M_{i,t}\right) + (1 - a)\left(M_{i,t-1} + T_{t-1}\right) \tag{6-1}$$

The coefficient α is the smoothing constant ($0 < a < 1$), which value is important for the forecast. The selection of the best value of α is based on the minimal sum of error squared.

$M_{i,t+1}$: The forecast of the M_i value for the next period.

$M_{i,t}$ **:** The calculated value of the M_i at time *t*.

$M_{i,t-1}$: The forecast for the M_i value of the previous period.

T_{t-1}: is the Trend estimate for the period t-1.

Before beginning a forecast we should set the initial values of $M_{i,0}$ = “first value of the PM” and $T_0 = 0$

The trend estimate for the next period is calculated by using *b* (as a second parameter):

$$T_t = b\left(M_{i,t} - M_{i,t-1}\right) + (1-b)\left(T_{t-1}\right) \tag{6-2}$$

Finally, the trend-adjusted forecast is denoted by

$$M_{i,t+k} = M_{i,t+1} + kT_t \tag{6-3}$$

where *k* indicates the periods that the forecasting is calculated, supposing that the same Trend will be valid for the future.

Hence, if M_i continues to exceed the specified threshold $Th_{M_i}^{C_j}$ in the time period $t+k$ ($M_{i,t+k} \xrightarrow{Th_{M_i}^{C_j}} C_j$) then the CNM is confident for addressing directly M_i and the associated CA C_j is candidate for enforcement.

$M_{i,t+k}$ is calculated, using equation (6-3), for each $M_i \in M$ that triggers a CA. According to $M_{i,t+k}$ value the lists of the triggered CAs (*C'*) and the respective associated PMs (*M'*) are updated. *C'* and *M'* are the inputs for the next step.

6.2.2 Conflicts on Configuration Actions

The scheme for the identification and the resolution of conflicts among the candidate CAs of *C'* is described in this sub-section. Three types of CAs could be identified. The SON coordination functionality detects a conflict if one of the following conditions is satisfied, during the same cycle of operation:

1. Metric $M_i', i=1$ requires the enabling of parameter C_j' (i.e., mode ON), while a second metric $M_i', i=2$ requires the disabling of parameter C_j' (i.e., mode OFF).
2. Metric $M_i', i=1$ requires the decrease of parameter C_j', while another metric $M_i', i=2$ requires the decrease of parameter C_j' e.g., AP transmission power.
3. Metric $M_i', i=1$ requires the minimization of parameter C_j', while another metric $M_i', i=2$ requires the maximization of the same parameter.

In the case that a conflict on $C_j' \in C'$, has been detected by the NECM or the NDCM, considering the above conditions. then only one direction for C_j' adaptation should be selected. The severity and the importance of the each M_j' that triggers the corresponding C_j' are assessed in order to select the appropriate direction. The priority of each M_j', as well its distance from the triggering threshold ($Th_{M_i'}^{C_j'}$) are the criteria for the assessment. For each M_j' the following function is calculated:

$$Q_{C_j',M_i'} = P(M_i')\left(M_i' - Th_{M_i'}^{C_j'}\right) \tag{6-4}$$

where $P(M_i')$ provides the priority value that has been assigned to M_j' according to the defined priority class. The PM M_j' that has the maximum $Q_{C_j',M_i'}$ value is identified and the corresponding direction/adaptation on the conflicting C_j' CA is selected.

6.2.3 Dependencies on Performance Metrics

After the identification of non-conflicting CAs in the context of each NECM and at the NDCM side the NECM checks whether the head of the cluster (i.e., NDCM) has identified a cluster-level problem. In that case the re-configuration that has been triggered by NDCM is prioritized and applied by the NECMs. Thereinafter, the problems that individual NECMs identify are solved. Firstly, the Impact Area Graph (IMAG) is build for the cluster area. The goal is to identify the highest priority problem in a 2 hops distance. Figure 6-3 depicts some characteristic examples of a IMAG for a simple network topology. We consider that according to the associated PMs the priority list starting from highest priority action is a follows: CA1, CA4, CA3, CA2. For instance, CA1 and CA2 that are triggered at AP4 and AP7 respectively (Figure 6-3-a), could be concurrently enforced regardless of their mutual priority. This is because there is not any direct or indirect impact on each other PMs. In a more complex scenario four CAs are triggered at AP1, AP4, AP6 and AP9 (Figure 6-3-b). In this case each AP selects the highest priority CA between itself and the neighboring NECMs. AP1 has to select between CA1 and CA2, AP3 selects between CA1, CA2, and CA3. Similarly AP6 selects between CA2, CA3 and CA4, while AP9 between CA3 and CA4. After this first selection phase an updated IMAG is formed, according to their common priority list. Finally, two actions are selected for a direct application: CA1, CA4. In the example that Figure 6-3-c illustrates CA1 is prioritized, by using the priority list presented above. Similarly, in Figure 6-3-d CA1 and CA4 are prioritized. The rest CAs are assessed in the next cycle of the NECM after the enforcement of prioritized CA. In the case that two CAs have the same priority then the distance from the triggering threshold $(Th_{M_i'}^{C_j'})$ is the criterion for the prioritization.

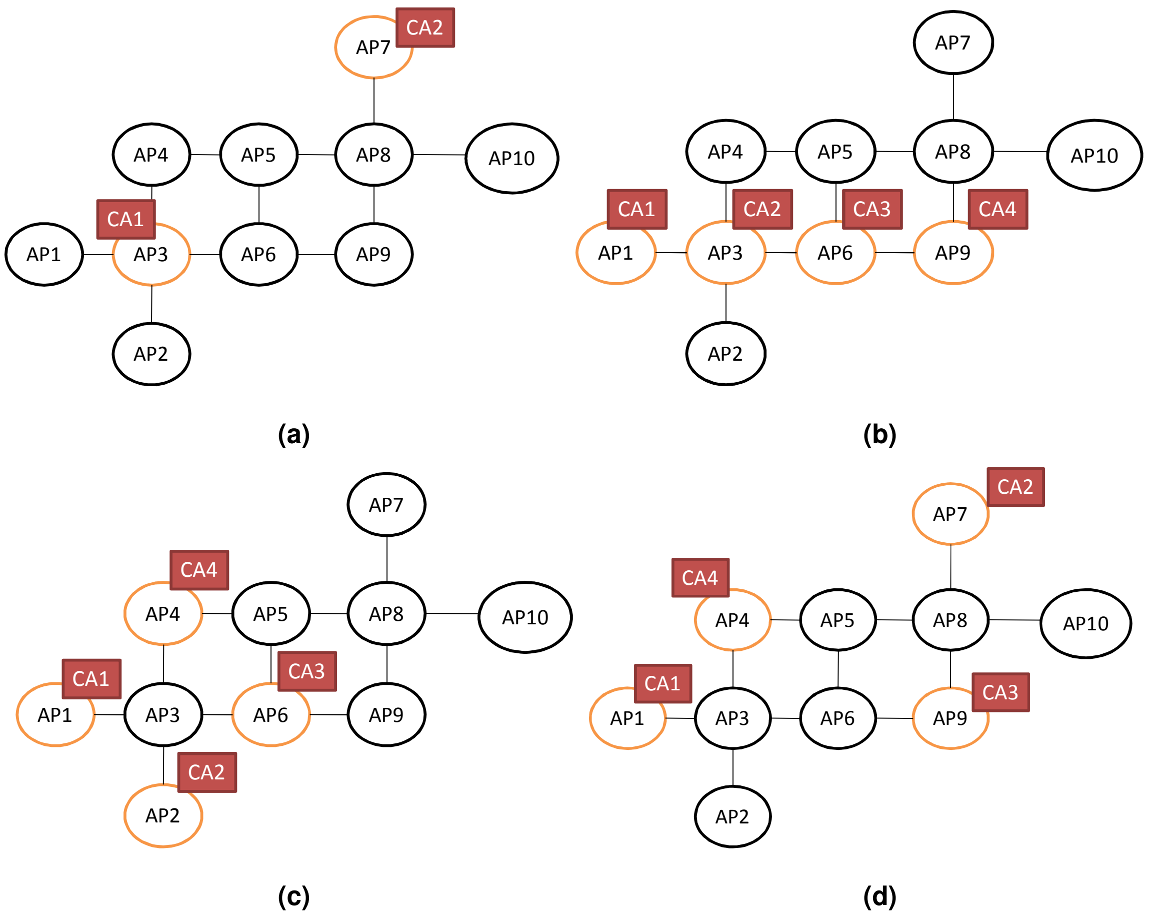


Figure 6-4: Impact Area Graph (IMAG) for a sample topology

After the selection of the appropriate action then the NECM estimates the benefit or the cost that is allocated on the rest of the performance metrics. The Cost-Benefit analysis can provide an estimation of the best alternative of the prioritized CA. This phase facilitates the decision making of NECM, by causing the least negative impact (or maximum positive effect) on the other PMs. At this point, we should note that costs (or benefits) are expressed via a different way for each PM. In the performance evaluation section we discuss CCO, energy saving and interference use cases.

6.2.4 Probabilistic Scheme for Dependencies on Performance Metrics

In the case that the direct Cost-Benefit analysis of a specific configuration action on a set of performance metrics is not feasible due to computational issues or other time constraints, then the adoption of a probabilistic scheme could be studied. In that case each PM needs to be described via a linguistic variable. Linguistic variables could reduce the overall computation complexity during their joint assessment. A numeric linguistic variable is characterized by $< \mu_k, L, r >$, where L is a finite set of linguistic terms $\{l_1, ...l_v\}$, which are assigned to μ_k according to a semantic rule r. This rule assigns each term $l \in L$ according to its value on μ_k. For instance, the Interference PM could be characterized as {low, medium, high}. An interference level of 0.6 using an example r rule is characterized as 'high'. Based on the linguistic variables, the CNM is aware about the identified states of each PM, as well as for the transitions that are beneficial (or not). Each PM μ_k consists states ($\sigma_s^{\mu_k}$) of PM.

The impact of each non-conflicting CA ($C_j'' \in C''$) on other PMs ($M_i \in M$) is assessed in order to select the most appropriate CA for execution. The goal is to identify C_j'' that has the highest priority and the less probability to lead either to conflicts on PMs or to trigger additional adaptations in the next cycle(s) of the respective CNM (NECM, NDCM).

Certainty Factor (CF) method [141] has been adopted in order to indicate CA that gives the higher confidence to network stability and total performance improvement. CFs are similar to conditional probabilities and they represent a measure of belief or disbelief in the outcome/result *h* of a rule that has the following form: "IF e THEN h"; instead of the degree of probability of the result. While probabilities range from 0 to 1, CFs range from -1 (believed not to be the case) to 1 (believed to be the case). For the estimation of the CF, it is necessary to calculate the Measure of Belief (MB) and the Measure of Disbelief (MD) as follows:

$$CF = \frac{MB(h,e) - MD(h,e)}{1 - \min\{MB(h,e), MD(h,e)\}} \tag{6-5}$$

This can lead to CF in the range -1.0 (total disbelief) to 1.0 (total belief). Or it can lead to CF's in the range 0.0 (no confidence) to 1.0 (total confidence) if MD is not measured and set to 0 in the formula. If MD is not used (as is often the case), then CF = MB.

$$MB(h,e) = \begin{cases} 1 & if\ P(h) = 1 \\ \max\left\{0, \dfrac{P(h|e) - P(h)}{1 - P(h)}\right\} \end{cases} \tag{6-6}$$

$$MD(h,e) = \begin{cases} 1 & if\ P(h) = 0 \\ \max\left\{0, \dfrac{P(h)\ -\ P(h|e)}{P(h)}\right\} \end{cases} \tag{6-7}$$

For SON problems, parameter *e* represents a CA ($C''_j \in C''$) that is under assessment, while *h* indicates a specific state transition of $M_i \in M$ (Figure 6-2-a) that is linked with $C''_j \in C''$. $M_{i,y} = \{M_{i,1}, M_{i,2}, ..., M_{i,l}\}$, $l \in \aleph$ denotes the identified states of each M_i. Please, note that various PMs do not have the same number of states. Considering that $M_{i,x}$ indicates the existing state of M_i, the goal of the CF is to estimate the MB and MD that the application of C_j will lead to a desirable state transition ($M_{i,x} \rightarrow M_{i,y}$) or that it will not lead to a state transition, which is not desirable for the network operation ($M_{i,x} \nrightarrow M_{i,y}$). In this case the rule of the “IF e THEN h” form is translated as follows:

IF “C''_j *is applied* THEN $M_{i,x} \rightarrow M_{i,y}$”

Equations (6-5) and (6-6) for MD and MB are adapted as follows:

$$MB(M''_{i,x} \rightarrow M''_{i,y}, C''_j) = \begin{cases} 1 & if\ P(M''_{i,x} \rightarrow M''_{i,y}) = 1 \\ \max\left\{0, \dfrac{P(M''_{i,x} \rightarrow M''_{i,y} | C''_j) - P(M''_{i,x} \rightarrow M''_{i,y})}{1 - P(M''_{i,x} \rightarrow M''_{i,y})}\right\} \end{cases} \tag{6-8}$$

$$MD(M_{i,x} \rightarrow M_{i,y}, C_j) = \begin{cases} 1 & if\ P(M_{i,x} \rightarrow M_{i,y}) = 0 \\ \max\left\{0, \dfrac{P(M_{i,x} \rightarrow M_{i,y})\ -\ P(M_{1,x} \rightarrow M_{1,y} | C_j)}{P(M_{i,x} \rightarrow M_{i,y})}\right\} \end{cases} \tag{6-9}$$

Exploiting the information of the bipartite graph (Figure 6-2-b), Table 6-1 describes the algorithm for the calculation of the CF for C''_j considering all possible states’ transitions of each PM that is linked with C''_j.

Table 6-1: Certainty Factor calculation

```
Calculate_Certainty_Factor ($C_j$, $M''$)

For each metric $M_i$ that is associated with $C_j$ do
    $M_{i,x} \rightarrow Current\ State\ Level;$
    For each possible state transition ($M_{i,y}$) of $M_i$ do
        If $M_{i,y}$ is a desirable transition then do
            $MB(M_{i,x} \rightarrow M_{i,y}, C_j) = \begin{cases} 1 & if\ P(M_{i,x} \rightarrow M_{i,y}) = 1 \\ \max\left\{0, \dfrac{P(M_{i,x} \rightarrow M_{i,y} \mid C_j) - P(M_{i,x} \rightarrow M_{i,y})}{1 - P(M_{i,x} \rightarrow M_{i,y})}\right\} & \end{cases}$
            $MD(M_{i,x} \rightarrow M_{i,y}, C_j) = \begin{cases} 1 & if\ P(M_{i,x} \rightarrow M_{i,y}) = 0 \\ \max\left\{0, \dfrac{P(M_{i,x} \rightarrow M_{i,y})\ -\ P(M_{1,x} \rightarrow M_{1,y} \mid C_j)}{P(M_{i,x} \rightarrow M_{i,y})}\right\} & \end{cases}$
            $CF_{S_{i,y}} = \dfrac{MB(M_{i,x} \rightarrow M_{i,y}, C_j) - MD(M_{i,x} \rightarrow M_{i,y}, C_j)}{1 - \min\{MB(M_{i,x} \rightarrow M_{i,y}, C_j), MD(M_{i,x} \rightarrow M_{i,y}, C_j)\}}$
        else
            $MD(M_{i,x} \nrightarrow M_{i,y}, C_j) = \begin{cases} 1 & if\ P(M_{i,x} \nrightarrow M_{i,y}) = 0 \\ \max\left\{0, \dfrac{P(M_{i,x} \nrightarrow M_{i,y})\ -\ P(M_{1,x} \nrightarrow M_{1,y} \mid C_j)}{P(M_{i,x} \nrightarrow M_{i,y})}\right\} & \end{cases}$
            $MB(M_{i,x} \nrightarrow M_{i,y}, C_j) = \begin{cases} 1 & if\ P(M_{i,x} \nrightarrow M_{i,y}) = 1 \\ \max\left\{0, \dfrac{P(M_{i,x} \nrightarrow M_{i,y} \mid C_j) - P(M_{i,x} \nrightarrow M_{i,y})}{1 - P(M_{i,x} \nrightarrow M_{i,y})}\right\} & \end{cases}$
            $CF_{S_{i,y}} = \dfrac{MB(M_{i,x} \nrightarrow M_{i,y}, C_j) - MD(M_{i,x} \nrightarrow M_{i,y}, C_j)}{1 - \min\{MB(M_{i,x} \nrightarrow M_{i,y}, C_j), MD(M_{i,x} \nrightarrow M_{i,y}, C_j)\}}$
        endIf
    endFor
    $CF_{j,i} = \dfrac{\sum_{p=1}^{m} CF_{S_{i,p}} + \sum_{q=1}^{n} CF_{S_{i,q}}}{l}\ ,\ m + n = l$
endFor
$\overline{CF_{C_j}} = \sum_{i=1}^{k} \left( P(M_i'') CF_{j,i} \right)$
```

The SON Coordinator calculates:

- the CF of the transition $M_{i,x} \to M_{i,y}$ in the case that C''_j is applied, where $M_{i,y}$ is a more desired state for the performance of the system juxtaposed to $M_{i,x}$.
- the CF for the avoidance of a transition $M_{k,x} \to M_{k,y}$, in the case that C''_j is applied, where $\mathrm{M}_{k,y}$ is not a desired state for the performance of the system.

The algorithm provides the system-wide CF for C''_j denoted $\overline{CF_{C''_j}}$, which considers the priority of each checked PM, denoted by $P(M''_i)$ and the CF of C''_j that each associated PM will have a positive transition. The algorithm is repeated for all candidate CAs ($C''_j \in C''$) and the SON coordinator selects to apply the CA that has the maximum $\overline{CF_{C''_j}}$ value. The SON Coordinator is aware about and stores the corresponding probabilities, conditional or not (e.g., $P(M_{i,x} \to M_{i,y})$, $P(M_{1,x} \to M_{1,y} \mid C_j)$).

6.3 Performance Evaluation

The mechanisms presented in section 6.2 have been implemented using the OPNET Modeler simulation software [143]. Performance results for the coordination of different SON use cases are described and discussed. The values of the simulation parameters are the same as in Table 5-5. We have chosen a network topology that consists of eight APs. The initial established network topology is presented in Figure 6-5. An edge exists between two APs, if one is within the coverage area of the other. Cognitive managers in this simulation environment address the operation of different types of performance improvement tasks (CCO, Interference, and Energy Saving). The performance metrics that are monitored are the following:

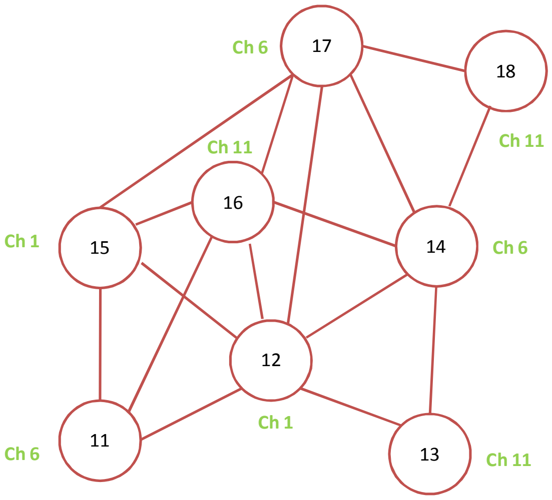


Figure 6-5: Access points simulation topology

- Coverage Optimization Opportunity (COOP),
- CUR: Capacity Usage Ratio (CUR),
- Access Point Energy Consumption (EC_{AP}^{Tx} , EC_{AP}^{Rx}),
- User Equipment Energy Consumption (EC_{UE}^{Tx} , EC_{UE}^{Rx}),
- BER.

The CAs that are supported by each NECM and NDCM are as follows:

- Load balancing,
- AP switch On,
- AP switch Off,
- UEs Relay Formation,
- Channel re-allocation.

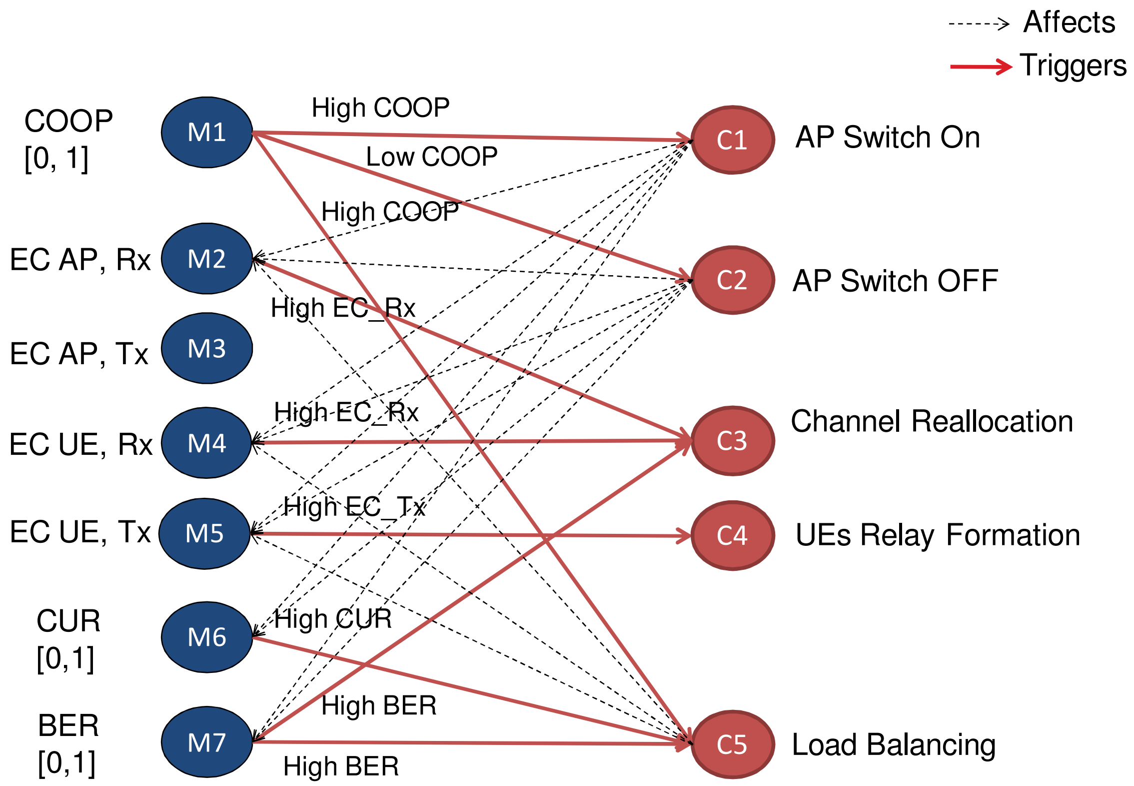


Figure 6-6: CCO, ES, Interference: PMs and CAs bipartite graph

Taking into consideration the analysis in section 5 for Energy saving and CCO the bipartite graph for the available PMs and CAs is depicted in Figure 6-6. The solid red line presents the CAs that each PM triggers if the predefined threshold has been exceeded, while the dashed lines present the PMs that might be affected after the enforcement of

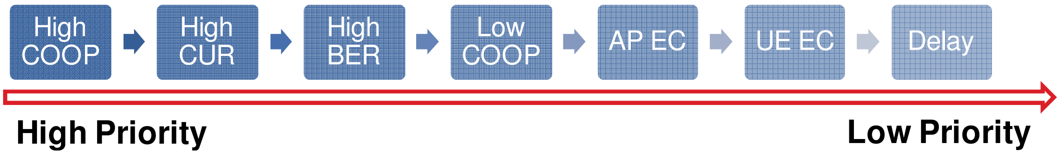


Figure 6-7: Performance metrics priority list

Table 6-2: NECMs: AP 11, AP 13, AP 14 and AP 15 monitoring

Performance Metric	AP11 (Channel 6)	AP13 (Channel 10)	AP15 (Channel 1)	AP14 (Channel 6)
AP DL Traffic (Mbps)	1.507	2.163	0.272	2.136
AP UL Traffic (Mbps)	0.234	0.024	0.468	0.000
AP Sensed Traffic (Mbps)	0.350	0.036	0.964	2.794
UEs Tx Traffic (Mbps)	0.346	0.027	0.696	0.000
UEs Sensed Traffic (Mbps)	1.495	1.719	1.756	4.891
CUR	0.265	0.313	0.138	0.492
OF	1.000	1.000	0.900	0.600
COOP	0.265	0.313	0.169	0.306
AP EC Tx (J/Sec)	0.805	1.155	0.145	1.141
AP EC Rx (J/Sec)	0.017	0.002	0.048	0.140
UEs EC Tx (J/Sec)	0.030	0.004	0.102	0.000
UEs EC Rx (J/Sec)	0.075	0.086	0.088	0.240
AP EC Tx (nJ/bit)	534.000	534.000	534.000	534.000
UE EC Tx (nJ/bit)	86.810	135.930	146.754	98.610
UEs BER	0.000037	0.000000	0.000028	0.000350

Table 6-3: NECMs: AP 12, AP 16, AP 17 and AP 19 monitoring

Performance Metric	AP12 (Channel 1)	AP16 (Channel 11)	AP17 (Channel 6)	AP18 (Channel 11)
AP DL Traffic (Mbps)	0.115	3.654	2.108	1.828
AP UL Traffic (Mbps)	0.099	1.168	0.450	0.447
AP Sensed Traffic (Mbps)	1.118	1.844	2.806	0.670
UEs Tx Traffic (Mbps)	0.148	1.876	0.659	0.656
UEs Sensed Traffic(Mbps)	2.317	19.856	8.570	4.223
CUR	0.038	0.790	0.395	0.355
OF	0.667	0.800	0.733	1.000
COOP	0.112	0.828	0.506	0.355
AP EC Tx (J/Sec)	0.062	0.951	1.126	0.976
AP EC Rx (J/Sec)	0.056	0.092	0.140	0.034
UEs EC Tx (J/Sec)	0.026	0.233	0.140	0.059
UEs EC Rx (J/Sec)	0.116	0.948	0.420	0.211
AP EC Tx (nJ/bit)	534.000	534.000	534.000	534.000
UE EC Tx (nJ/bit)	179.315	124.305	211.938	90.381
UEs BER	0.000105	0.000023	0.000219	0.000001

a CA. As it is described in section 6.2, the coordination of various SON problems in the context of the same or different cognitive managers, requires the formation of PMs' priority list. Figure 6-7 illustrates the priority list that has adopted for this experimentation scenario.

The initial traffic levels and the values of all parameters that are monitored by each NECM (i.e. AP NECM) are presented in Table 6-2 and Table 6-3. We observe that AP16 is too loaded having a high COOP value (>0.8). In addition, AP 16 and AP 17 have high BER (>10^{-4}), due to the overlap of the frequency channels that they use. AP 12 has low UL and DL traffic and many neighbouring APs; hence a low COOP value is calculated (<0.15). All the other PMs of APs are within the acceptable thresholds. Figure 6-8 illustrates the problems that are detected per AP. Thus, different configuration actions are concurrently triggered: a) AP 16 calls for a load balancing action, b) AP12 triggers its switch-off action, c) AP 17 and AP 14 trigger a channel re-allocation action to address the high (co-channel) interference status. At the cluster level there is not any detected problem, since all PMs are at the appropriate state (Table 6-4).

In this case the proposed SON coordination scheme is applied. Firstly, the IMAG is formed, taking into account that all the detected problems are not the result of a temporary change, while there is not any cluster-level problem (Table 6-4). Using the priority list of Figure 6-7 the first CA that is enforced is the balancing of AP 16 load. Based on the Cost-Benefit analysis and considering that the next PM in the priority list is BER, the half of the traffic of AP16 is served via AP15. We have used the load balancing scheme that has been presented in section 5.4.1. Then AP14 changes the selected frequency channel. Channel 3 is selected by AP 14 and the associated terminals, using

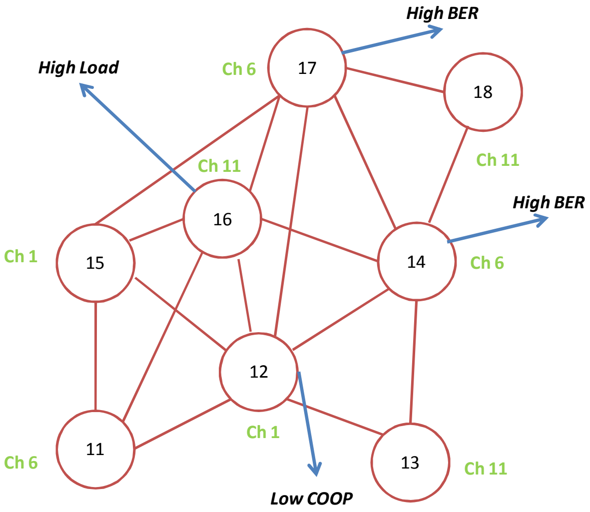


Figure 6-8: Identified problems in the simulation topology

Table 6-4: NDCM (cluster level) monitoring before self-organization

Performance Metric	Value
AP DL Traffic (Mbps)	13.834
UEs Tx Traffic (Mbps)	4.390
CUR	0.325
OF	0.571
COOP	0.527
AP EC Tx (J/Sec)	0.923
AP EC Rx (J/Sec)	0.066
UEs EC Tx (J/Sec)	0.074
UEs EC Rx (J/Sec)	0.273
AP EC Tx (nJ/bit)	534.000
UE EC Tx (nJ/bit)	135.066
AP EC Rx (nJ/bit)	29.033
UE EC Rx (nJ/bit)	119.645
UEs BER	0.0000910

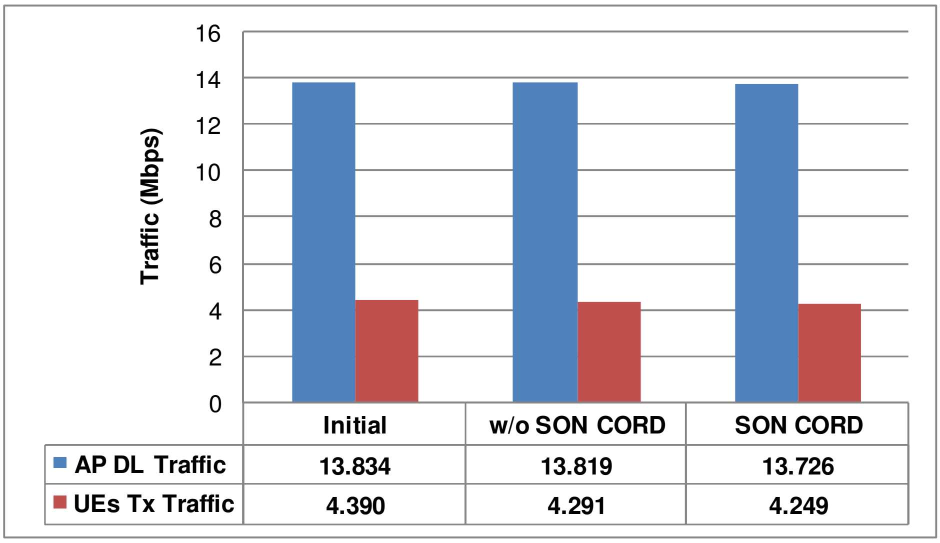
the scheme presented in section 3.5.3.2, in order to reduce the BER of AP 17 and AP 14. Finally, since the COOP metric of AP 12 is still below B_d threshold, then AP 12 is switched off and the terminals are handed over to AP 15 (section 5.4.1).

The SON coordination scheme allows the performance improvement assessing the severity of detected problems, without proceeding to needless adaptations that affect the stability of the network and extend its transitory period. Figure 6-9, Figure 6-10, Figure 6-11 and Figure 6-12 illustrate the change of cluster-level traffic, BER, AP energy consumption and UE energy consumption after the different actions that the SON

Table 6-5: NDCM (cluster level) monitoring after SON Coordination

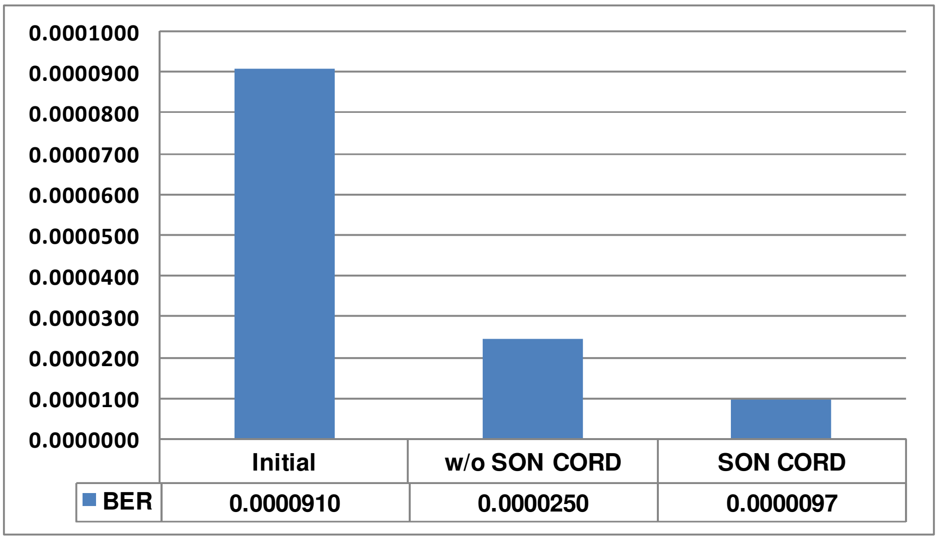
Performance Metric	Value
AP DL Traffic (Mbps)	13.726
UEs Tx Traffic (Mbps)	4.249
CUR	0.367
OF	0.476
COOP	0.620
AP EC Tx (J/Sec)	0.922
AP EC Rx (J/Sec)	0.031
UEs EC Tx (J/Sec)	0.116
UEs EC Rx (J/Sec)	0.285
AP EC Tx (nJ/bit)	534.000
UE EC Tx (nJ/bit)	191.882
AP EC Rx (nJ/bit)	12.060
UE EC Rx (nJ/bit)	111.038
UEs BER	0.0000097

coordination process has prioritized. Furthermore, the values of the above-mentioned metrics are also provided for the case that the SON coordination mechanism is not enabled. Considering a rational scenario, the load of AP 16 is handed over to AP 12 and AP 15. AP12 COOP is increased and the respective switch off action is not triggered. Furthermore, AP 14 selects channel 3 to reduce the detected BER. Although the behaviour of the network managers is rational as regards the sequence of the adaptations, we notice that BER and the network side energy consumption have not reached the level of the experiment where SON coordination is enabled. On the other hand the average energy consumption of user equipment is increased due to the longer transmission distance, after load balancing and the switch off actions. The formation of the priority list is important for the behaviour of a self-organizing network, especially in the case that a SON coordination scheme is used.



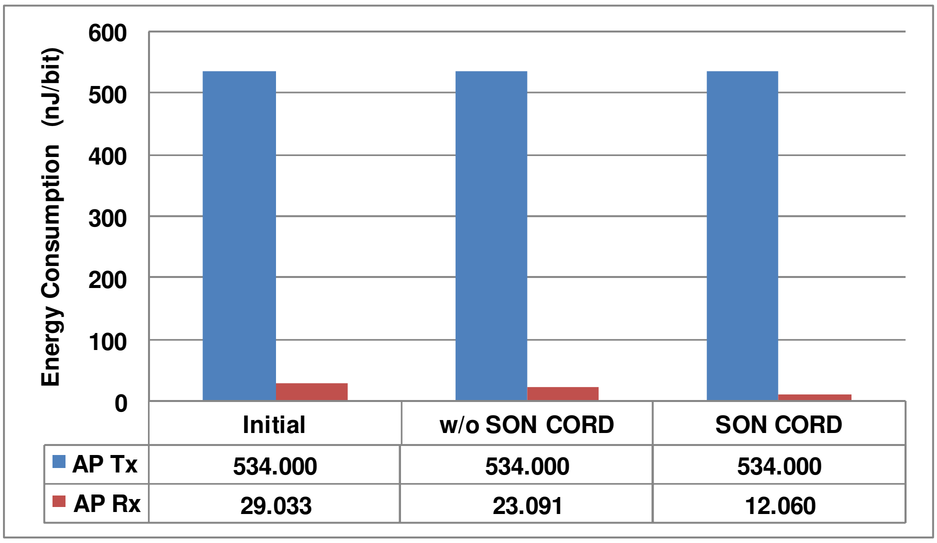
	Initial	w/o SON CORD	SON CORD
🟦 AP DL Traffic	13.834	13.819	13.726
🟥 UEs Tx Traffic	4.390	4.291	4.249

Figure 6-9: Cluster-level traffic (UL/DL)



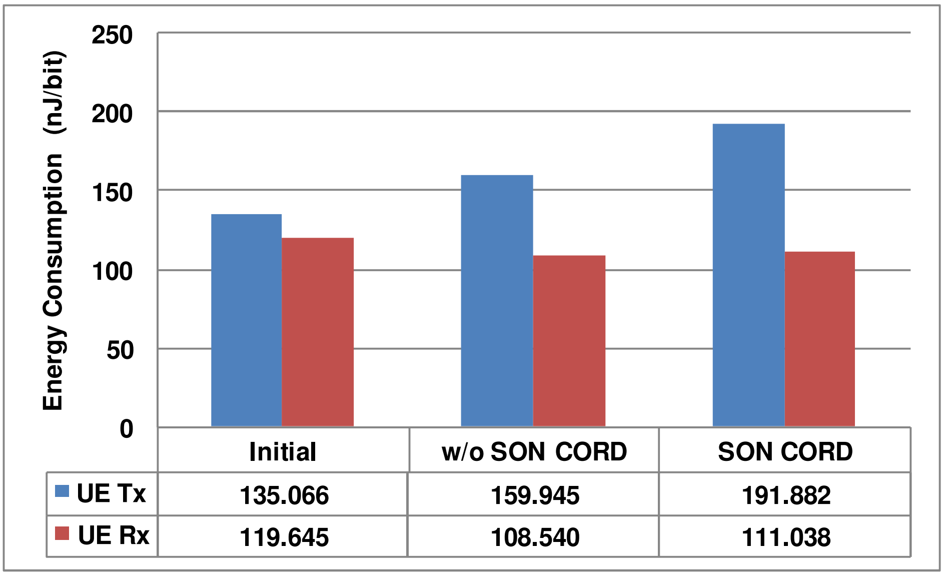
	Initial	w/o SON CORD	SON CORD
🟦 BER	0.0000910	0.0000250	0.0000097

Figure 6-10: Cluster-level BER



	Initial	w/o SON CORD	SON CORD
🟦 AP Tx	534.000	534.000	534.000
🟥 AP Rx	29.033	23.091	12.060

Figure 6-11: Cluster-level nJ/bit AP energy consumption



	Initial	w/o SON CORD	SON CORD
🟦 UE Tx	135.066	159.945	191.882
🟥 UE Rx	119.645	108.540	111.038

Figure 6-12: Cluster-level nJ/bit UE energy consumption

CHAPTER 7

SERVICE-AWARE COGNITIVE SELF-MANAGEMENT OF WIRELESS NETWORKS

In this chapter we describe an algorithmic framework for the incorporation of cognitive capabilities in network management, facilitating performance management tasks. We take into consideration the system architecture that has been described in chapter 4. This framework is used for the improvement of VoIP QoS in a congested WiMAX network. Despite the WiMAX related introduction in the rest of this chapter the proposed solution is not access technology specific, but is equally feasible to other broadband wireless networks as well, such as WLAN and LTE.

The proposed algorithmic framework consists of the decision making, the execution and the learning phase [146]. The decision making part includes the scheme for the identification of the most appropriate action for the packet loss reduction of VoIP service; selecting between a) the change of VoIP flows priority at the base station, exploiting MAC features, and b) the change of VoIP flows selected codec, exploiting service level features. The solution and the quantification of the derived action (e.g., number of VoIP flows, the type of codec transition) is achieved using either a history-based scheme that takes advantage of previous events, or a heuristic approach for un-classified (i.e., unknown) situations. A k-Means learning algorithm is introduced to process the accumulated knowledge from all applied actions and evolve the decision making scheme [147], [148]. The performance and feasibility evaluation of the proposed solution has been tested using FIRE Panlab WiMAX experimental facility [149].

VoIP constitutes one of the most flourishing services for future networks [150]. As it is presented below, several effective techniques have been proposed in the literature, for mitigating the traffic congestion of a VoIP service in a WiMAX network e.g., [151], [152], [153], [154]. But most of them are based on predetermined problem-solving schemes, without evolving their decision making during their lifetime, by exploiting the feedback from performed actions or historic data. Moreover, existing works are mainly focusing on an individual re-configuration action, without assessing the different types of adaptations that could be used.

Central to our approach is the cooperation between the network self-management system and the service stratum [155].

7.1 Mapping of System Capabilities to a Logical Network

The service stratum includes functions that control and manage network services to enable the end-users services and applications [156]. Both service-level monitoring data and service-level adaptation actions are useful to a NMS entity (NECM, NDCM) for efficient network adaptation and QoS assurance. For that purpose, two cooperation channels (i.e., interfaces) are identified from the service stratum to the NMS, enabling service-aware network self-management. The first one is used for the provision of monitoring data (e.g., service type, VoIP codec, data rate) from the service stratum to the network management level. The second cooperation channel is used by the NMS in order to trigger a service-level adaptation action. Thus, NECM and NDCM extend their list of available configuration actions, for the recovery of a detected fault or for communication network performance improvement.

In this case, the NECM is deployed to a WiMAX base BS for its local management tasks, while the NDCM is deployed at the network side, being also responsible for the interactions between NMS and the service domain (Figure 7-1). The allocation of the cognitive cycle phases (monitoring, decision making, execution, and learning) at NECM, NDCM and service stratum entities is depicted in Figure 7-2. The NECM of the BS constantly monitors network device statistics (e.g., UL/DL, TCP, User Datagram Protocol (UDP) parameters), which are periodically transmitted to the corresponding NDCM. The service stratum periodically monitors the perceived QoS (e.g., packet loss) and the type of service that each client consumes, as well as service profile information from the service providers (e.g., VoIP Codec).

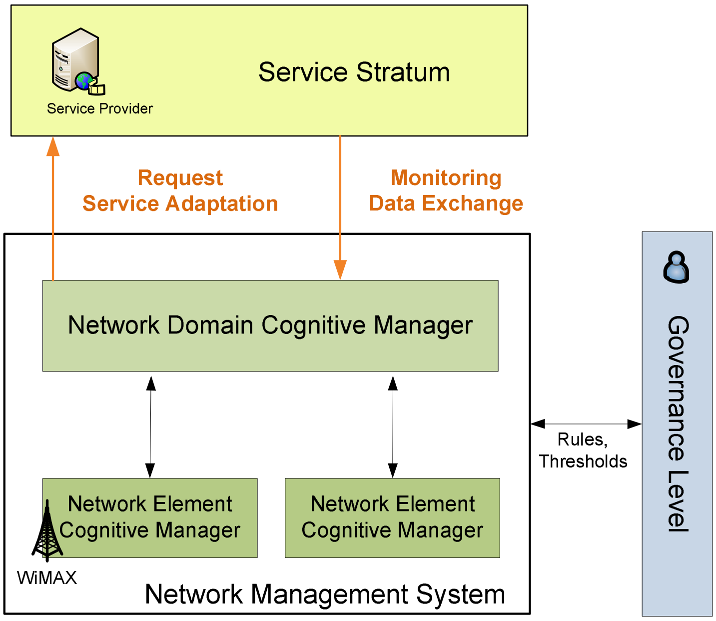


Figure 7-1: Network Management and service stratums interaction

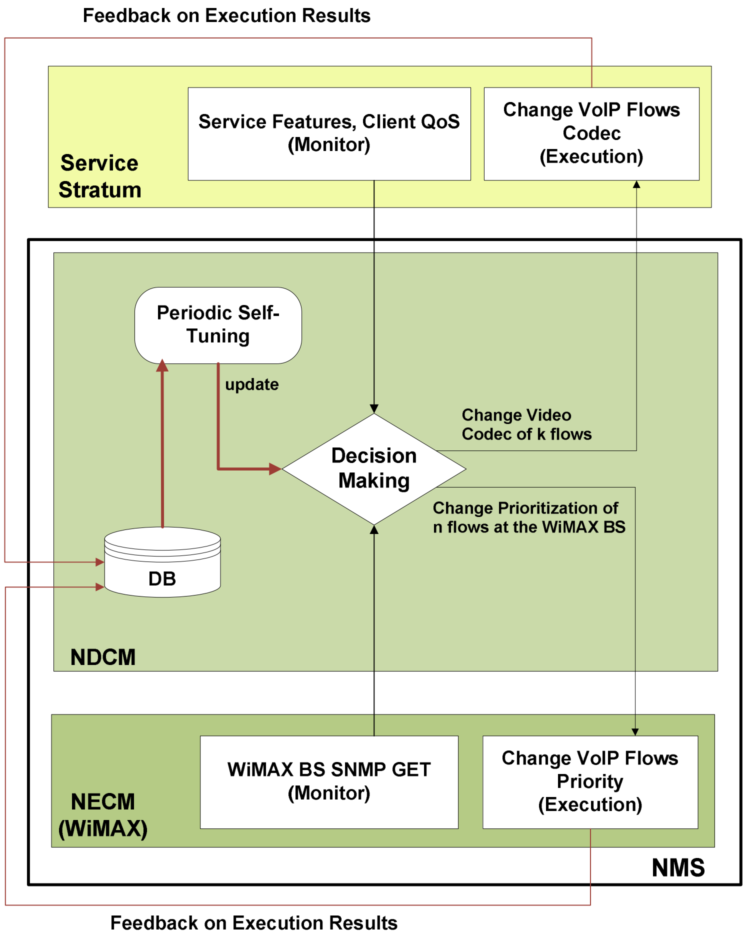


Figure 7-2: Algorithmic framework activity diagram

The decision making engine of the NDCM filters the collected monitoring data from the NECM and the service stratum in order to trigger a performance improvement action, according to the specified rules and QoS requirements. In this case, the NDCM decision making engine identifies high average Packet Loss (PL) values for the users that consume a VoIP service and decides on the most appropriate action selecting between a network-side (i.e., change VoIP flows priority) or service-side adaptation (i.e., change codec of VoIP flows). The outcome of the decision making phase is transferred either to the BS NECM or to the service stratum for its enforcement.

The NDCM collects the performance outcome of all the applied actions and processing periodically the accumulated feedback, via learning techniques, fine tunes thresholds and parameters of the decision making scheme.

7.2 Cognitive Decision Making for VoIP QoS Assurance in a IEEE 802.16 Network

The goal of the decision making phase at the NDCM is to reduce a detected high PL level (PL_d), below the acceptable threshold (PL_{thr}), which is set by the network administrator. Two adaptation actions are studied:

- Change of the priority of VoIP flows at the WiMAX BS side from normal to high priority class (network- side adaptation).
- Change of the codec of VoIP flows considering five different VoIP Codec types (service-side adaptation).

The proposed decision making scheme initially selects the appropriate action and then identifies the intensity of the selected adaptation (e.g., number of adapted VoIP flows, codec transition). The latter is necessary, since a balance is required between PL reduction and other quality or service delivery requirements. For instance, the change of all flows to the less demanding codec will reduce the PL to the lowest possible level. However, it also reduces the data rate of all VoIP flows, leading to the degradation of end users' service experience. Two methods have been proposed for calculating the intensity of the selected adaptation: a) a heuristic approach for un-classified situations, and b) a history-based method that exploits the accumulated knowledge from previous events. The steps of the decision making scheme are illustrated in Figure 7-3.

The total number of VoIP flows that traverse the WiMAX BS, denoted by N, consists of the following type of flows, according to the used codec: N_1 of G.711.1 codec, N_2 of G.711.2 codec, N_3 of G.729.2 codec, N_4 of G.729.3 codec, N_5 of G.723.1 codec.

$$N_1 + N_2 + N_3 + N_4 + N_5 = N \tag{7-1}$$

Each codec type has a different bit rate (kbps) and consequently a different impact on PL_d. The VoIP flows that use G.711.1 codec have the maximum nominal bit rate (64 kbps), while G.723.1 VoIP flows have the minimum nominal bit rate (5 kbps). Let X_i denote the bit rate of codec i , and X_1 the maximum bit rate of codec G.711.1. Hence, the impact of each codec on PL_d is expressed as follows:

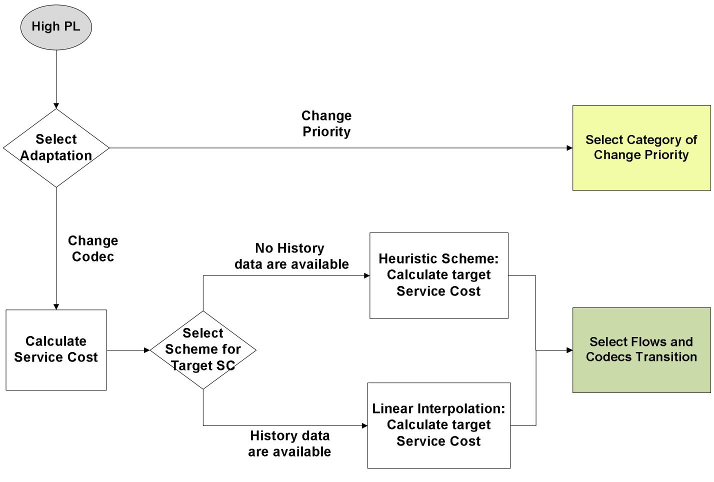


Figure 7-3: Decision making scheme for adaptation selection

$$W_i = \frac{X_i}{X_1}, i = 1...5 \tag{7-2}$$

Apart from VoIP flows' traffic, each BS is also traversed by other type of services, which inject additional TCP and UDP traffic in the network. This traffic is named as background traffic (BT), which also affects PL_d.

7.2.1 Adaptation Action Selection

The first step of the algorithm includes the selection of the most appropriate adaptation action (Figure 7-3). The distance of PL_d from PL_{thr} is the main parameter for this decision. We should note that PL is measured in percentage terms (%). As it is shown in the experimentation section, each adaptation action has a different impact on PL_d reduction. Specifically, a slight reduction of PL_d could be achieved through VoIP flows' change priority at the BS side. On the other hand, the modification of the selected codec is more appropriate for a high PL event. For the selection of the adaptation type, the NDCM firstly checks whether the change priority of VoIP flows is sufficient for reaching PL_{thr}. If the latter is not achieved, then the adaptation of VoIP flows' codec is selected. In the case that all VoIP flows' priority is already set to high, then the NDCM proceeds directly to the change codec action.

We introduce the PL_{max} parameter to describe the maximum PL improvement that could be achieved via VoIP flows' change, from a normal to a high priority class. $PL_{\max}$ is affected by the VoIP traffic that is injected in the network domain, denoted by T_{VoIP}, and the monitored TCP or UDP BT level that traverse the WiMAX BS, denoted by T_{BT} (e.g., TCP = 4Mbps and UDP = 3Mbps). Hence, we also introduce factor I, which is a smoothing parameter that we use in order to tune PL_{max}, taking into account the VoIP traffic level (VR) in conjunction with the CUR of the WiMAX BS. Factor I is calculated as follows:

$$I = CUR \cdot VR \tag{7-3}$$

The *CUR* metric is defined as the fraction of the available capacity that is actually being used:

$$CUR = \frac{T_{VoIP} + T_{BT}}{C_{WiMAX}} \tag{7-4}$$

where C_{WiMAX} denotes the maximum available capacity of the WiMAX BS. *VR* parameter is calculated as follows:

$$VR = \frac{T_{VoIP}}{C_{WiMAX}} \tag{7-5}$$

Hence, Using PL_{max} we can calculate the expected level of PL improvement, denoted by PL_g, which could be achieved by applying the change priority of all VoIP flows:

$$PL_g = \begin{Bmatrix} f(PL_d) - \left(f(PL_{\max})\right)^I, & if\ T_{BT} > 0 \\ PL_d, & if\ T_{BT} = 0 \end{Bmatrix} \tag{7-6}$$

where function *f* converts PL from [0, 1] (i.e., percentage) to [0, 100] scale.

After the calculation of PL_g, using equation (7-6), we select the appropriate adaptation action, as follows:

- If PL_g is lower than PL_{thr}, the change of VoIP flows' priority is considered beneficiary for the network, reducing PL_d below PL_{thr}. The corresponding method is presented in Section 7.2.2.
- If PL_g is equal or higher than PL_{thr} then the change priority action might be ineffective for PL reduction; thus, the change of VoIP flows' codec is selected, which is described in Section 7.2.3.

7.2.2 Change of VoIP Flows Priority at IEEE 802.16 BS

The intensity of the change of VoIP flows' priority for PL reduction depends on the VoIP level that traverses the network and the used codec type. Hence, in order to avoid redundant adaptations, we have identified three categories, regarding the number of the VoIP flows that could be moved from a normal to a high priority queue at the WiMAX BS:

- Category I: all VoIP flows are adapted.
- Category II: half of the VoIP flows are adapted, starting from those that use the most demanding codec.
- Category III: one third of VoIP flows are adapted, starting from those with the most demanding codec.

For the selection of the appropriate category, we extend equation (7-6) by including parameter t_k, which indicates the impact of the selected category on the PL improvement (i.e., involved number of VoIP flows):

$$PL_g = PL_d - t_k (PL_{\max})^I \tag{7-7}$$

The value of t_k is different for each category and the initial value could be set as follows: $t_k = 1$ for category I, $t_k = 1/2$ for category II, and $t_k = 1/3$ for category III. At this point, we should note that the t_k value for each category could be dynamically adapted, using the proposed self-tuning method.

7.2.3 Change of VoIP Flows Codec

For the change of VoIP flows' codec type, it is needed to calculate the Service Cost of the VoIP flows, denoted by SC_d for the corresponding PL_d. SC_d provides an indication of the relative cost that is introduced in the network by the VoIP service. The maximum SC_d value, for a specific number of VoIP flows, appears if all flows use the most demanding VoIP codec (G.711.1), while the minimum SC_d value is calculated if all VoIP flows use G.723.1 codec (i.e., low data rate). Taking into consideration the number of the flows for each codec as well as codec's weight, presented in (7-1) and (7-2) respectively, we calculate SC_d as follows:

$$SC_d = \sum_{i=5}^{5} \left(N_i W_i \right) \tag{7-8}$$

Consequently, for PL_d reduction to the PL_{thr} level, through the change of VoIP flows' codec we have to estimate the target service cost, denoted by SC_g. We propose two algorithms for SC_g calculation:

1) A heuristic scheme, when there is not enough knowledge from previous VoIP codec adaptations.
2) A history-based algorithm, where the NDCM processes the collected historic data from previous adaptations through a linear interpolation scheme.

The history-based approach is used by the NDCM, if the following conditions are met:

- There are at least *m* different PL measurements (e.g., *m=10*) and
- the background traffic of these *m* measurements, is in the same BT range, and specifically $T_{BT} \pm a$ (7-9)

where parameter *a*, which is set by the network administrator, defines an interval for the classification of PL measurements, facilitating the accurate operation of the history-based approach. We have included the $T_{BT} \pm a$ condition, because the BT level in the network affects the correlation between SC_d and PL_d, and consequently the interpretation of the information that service cost provides. The above-mentioned heuristic and history-based schemes for SC_g calculation are presented below.

7.2.3.1 Heuristic algorithm for service cost calculation

In the heuristic scheme the NDCM does not have any previous experience for the change of VoIP flows' codec in the specific BT level. Hence, we introduce metric SCP_d to correlate PL_d with the calculated SC_d:

$$SCP_d = SC_d^{PL_d} \tag{7-10}$$

SCP_d actually correlates the network capacity or resources usage for a specific number of flows and the perceived QoS, as the latter is expressed by the detected PL. The optimal value of SCP_d equals to one, which means that there is no detected PL. The range of SCP_d is $\left[1, SC_N^{\max}\right]$, where $SC_N^{\max}$ denotes that all VoIP flows use the most bandwidth demanding codec.

The NDCM is aware of the upper bound of the acceptable PL level ($PL_{thr} < PL_d$) and estimates the target SCP value (SCP_g), after the enforcement of a successful change codec adaptation, considering that in the worst case it will be equal to SCP_d. Given that SCP_g $(SCP_g = SCP_d)$ and PL_{thr} values are known we can calculate SC_g solving (7-11).

$$SCP_g = SC_g^{PL_{thr}} \tag{7-11}$$

7.2.3.2 History-based algorithm for service cost calculation

As it is describe above, this scheme is utilized in the case that the decision making engine is aware of the effectiveness of previous codec adaptations, for the specific BT level. For that purpose the linear interpolation is used, which is a method for curve fitting using linear polynomials. It is widely used for tackling with numerical analysis problems. Considering two known points (x_0, y_0) and (x_1, y_1), the linear interpolant is the straight line *ε* between these points.

For a value x in the interval (x_0, x_1), the value y along ε can be given from the following equation:

$$\frac{(x - x_0)}{(y - y_0)} = \frac{(y_1 - y_0)}{(x_1 - x_0)} \tag{7-12}$$

Solving (7-12) for y gives:

$$y = y_0 + \left(x - x_0\right)\frac{\left(y_1 - y_0\right)}{\left(x_1 - x_0\right)} \tag{7-13}$$

A straightforward approach for using the aforementioned mathematical method to the reference problem of this work is to map *PL* to *x* and service cost (S*C*) to y, so as to develop an interrelation between the key factors of our decision making scheme. Towards this direction and considering a specific service scenario with known PL_d ($PL_d \succ PL_{thr}$) and SC_d, the coordinates (x_1, y_1) can be mapped to (PL_d, SC_d). Similarly, the other set of coordinates (x_0, y_0) can be retrieved by the history base of the mechanism selecting a relevant service with the same number of flows $N_i \in [1, 5]$ on the one hand and a low value of $PL_{History}$ ($PL_{History} \succ PL_{thr}$) on the other hand. Following this rational, (x_0, y_0) should be mapped to ($PL_{History}, SC_{History}$) The last step of the linear interpolation is to define a point *x* in the interval (x_0, x_1) so as to solve (7-13) for *y*. Obviously, the appropriate point *x* is PL_{thr} and the result of the linear interpolant provides the appropriate service cost SC_g in order to achieve the target value of PL_{thr} .

7.2.3.3 Scheme for the selection of VoIP flows codec transition

After the calculation of SC_g , using either the heuristic or the history-based method, the NDCM selects the VoIP flows that should be adapted as well as the necessary codec transitions to achieve a higher data compression. The latter reduces SC_d to the estimated SC_g level. For *N* VoIP flows that traverse the WiMAX BS and for *n* types of codec, the total number of all possible combinations for the allocation of the candidate codecs to the *N* flows is calculated as follows [157]:

$$C\left(N + n - 1, N\right) = \frac{\left(N + n - 1\right)!}{N!\left(n - 1\right)!} \tag{7-14}$$

For instance, if we have *k=7* VoIP flows and *n=5* codecs, there are 330 possible combinations of codecs for *k=7* flows. Thus, for a large number of VoIP flows there is also a very large set of combinations of codecs for the available flows. It is necessary to reduce the number of candidate combinations, by selecting those that will lead us efficiently below PL_{thr}.

Ten possible transitions have been identified among the five types of VoIP codec. Each codec change has a transition cost C_s, which depends on the data rate degradation that takes place after each transition (Figure 7-4):

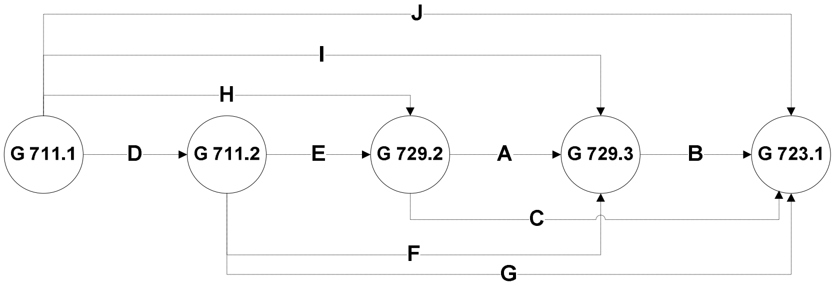


Figure 7-4: VoIP codec possible transitions

$$C_A \prec C_B \prec C_C \prec C_D \prec C_E \prec C_F \prec C_G \prec C_H \prec C_I \prec C_J$$

We denote as $F \subset \{f_1, f_2, ..., f_i\}$ the list of VoIP flows and $G \subset \{G_1, G_2, ..., G_i\}$ the selected codec per flow. Table 7-1 presents the proposed method for the calculation of the necessary codec transitions, starting from those VoIP flows that are less disruptive in terms of a codec transition (i.e., smaller data rate VoIP flows are prioritised).

Table 7-1: Codec Transition: Less disruptive first

```
Less_Disruptive_Codec_Transition ($SC_d$, $SC_g$, $C_i$, $F$, $G$)
SC_TMP = SC_d;
Do
    $\{C_s, f_i\}$ = Find the less costly VoIP codec transition;
    G → update selected Codec of VoIP flow $f_i$;
    C_TMP = C_TMP – C_s
while (SC_g < SC_TMP)
return {F, G}
```

7.3 Feedback-based Learning for Self-Tuning of Decision Making

The automatic fine-tuning of the decision making scheme is a key requirement for the effectiveness of a cognitive manager. This learning phase updates thresholds or parameters that are used by the decision making algorithms. The administrator of the cognitive managers sets the initial values of these parameters or thresholds, based on his accumulated experience. Even if their initial values are correctly set, the evolution of network status might call for their update in order to enhance the decision making. The cognitive manager should have the capability for on-line tuning of network thresholds/parameters, exploiting the feedback from previous actions and historic data.

The proposed learning method focuses on the update of the maximum PL improvement threshold (PL_{max}) that could be achieved via VoIP flows' priority change. This parameter is important for our algorithmic framework, since it provides the pilot for the selection of the appropriate action (Section 7.2). The NDCM stores the PL_d values, before the application of a reconfiguration action, as well as the PL_d value after the application of the change priority. Hence, the NDCM using its database (i.e., the history of previous actions) periodically checks whether the increase or the decrease of the PL_{max} is required for the effectiveness of the decision making.

The proposed learning method is based on the high dimensional Euclidean geometry of the k-Means algorithm [158]. k-Means is a fast and efficient clustering algorithm extensively employed in the area of unsupervised learning. The core idea is to partition a set of *N*, d-dimensional, observations into *k* groups such that intra-group observations exhibit minimum distances from each other, while inter-group distances are maximized.

Firstly, the NDCM retrieves the detected PL values that have triggered the change priority action and using the k-Means algorithm builds two clusters: *L1* and *H1* (Figure 7-5). Cluster *L1* includes those PL values that are labeled as low, while *H1* includes those PL values that are labeled as (relatively) high. At this point, we should clarify that only PL values below PL_{max} are retrieved and processed, since the values higher than PL_{max} have not triggered the change priority action. Thereinafter, the NDCM retrieves the sample of

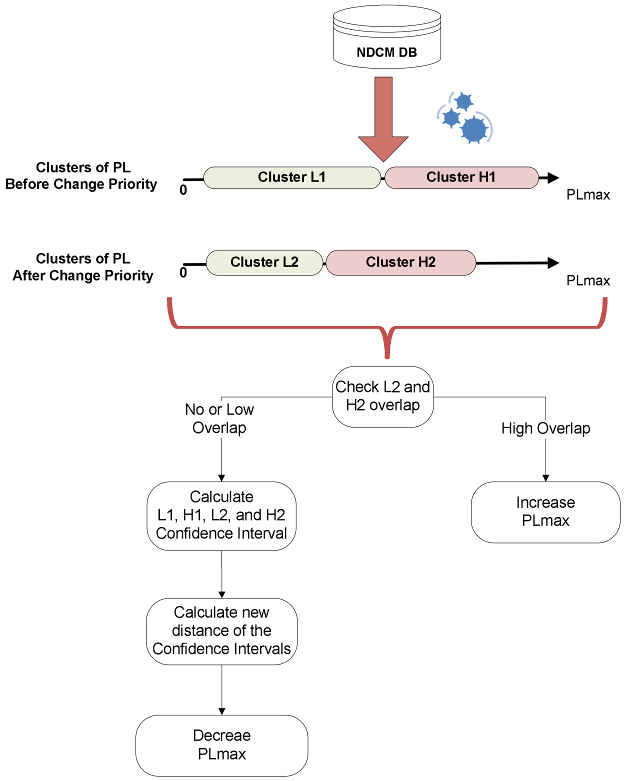


Figure 7-5: Self-tuning scheme for decision making thresholds

detected PL values after the enforcement of priority change for all VoIP flows. The k-Means algorithm is again used in order to build these two clusters that classify the PL values into a Low PL set (*L2*) and a High PL value set (*H2*). The centroids of clusters *L2* and *H2* are normally shifted left, comparing to the centroids of the clusters *L1* and *H1*. This implies that the PL has been decreased after the adaptation enforcement. The question is whether the PL rate has been decreased below the PL_{thr}. Based on this observation our method decides to update (increase or decrease) the PL_{max} value.

The increase of PL_{max} is deemed as the appropriate update if we have noticed that the change of the VoIP flows' priority from normal to high has been very effective. This is because it is less disruptive for the end-user perceived QoS to change VoIP flows' priority at the BS, instead of the VoIP codec. The PL_{max} increase is deduced, based on two conditions: a) if the centroids of the two clusters are too close or b) if there is a great overlap between clusters *L2* and *H2*, while both are lower than the PL_{thr} (Figure 7-5). This means that high PL values of *H1* cluster have been adequately reduced. In that case we increase PL_{max}, and consequently the range where priority change could be applied. PL_{max} is updated as follows:

$$PL_{\max} = PL_{\max} + b \tag{7-15}$$

where parameter *b* denotes the factor of the increase. The initial value of *b* is too small (*b*=0.005), which is gradually and carefully increased in order to avoid system instability from ping-pong effects (i.e., back and forth adjustments of PL_{max}). The increase of *b* value could be proportional to the overlap of *L2* and *H2* clusters.

On the other hand, if the NDCM notices that the change priority action is not effective for the PL reduction then the PL_{max} parameter should be decreased. The ineffectiveness of change priority is detected if PL values of *H2* cluster are above PL_{thr}. For the estimation of the PL_{max} decrease we calculate the confidence interval (CI) and then the distance between clusters *H1* and *H2*. The CI of each cluster (*H1*, *H2*) gives an estimated range of values from a given set of sample data. We have selected CI, instead of the centroid or the upper bound of each cluster in order to avoid noise data. The CI is calculated for specific confidence limits, which are the lower (H^{lb}) and upper (H^{ub}) boundaries/values of a confidence interval. These values define the range of a confidence interval. We have selected 90% confidence limit. For example, a *100(1-b)%* confidence interval for the mean of a normal population is:

$$\left(\mu - \frac{Z_{\alpha/2}\sigma}{\sqrt{N}}, \mu + \frac{Z_{\alpha/2}\sigma}{\sqrt{N}}\right) \tag{7-16}$$

Where μ is the sample mean, $Z_{a/2}$ is the upper $a/2$ critical value of the standard normal distribution, which is found in the table of the standard normal distribution, σ is the known population standard deviation, and *N* is the cluster size.

Hence, calculating the confidence interval of clusters *H1* (H_1^{lb}, H_1^{ub}) and *H2* (H_2^{lb}, H_2^{ub}) we can derive the updated PL_{max}, which is equal to the distance between the upper CI bounds of clusters *H1* and *H2*.

The proposed learning scheme could be also used to self-tune the rest parameters of the decision making algorithm. For instance, parameter t_k in equation (7-7) and parameter *a* in equation (7-9), which both initial values are set by the network administrator for specific network conditions.

7.4 Experimental Facilities Description

In this section, we present the configuration of FIRE Panlab and CORE experimental facilities that have been used to test the proposed algorithmic framework. The testing facility connecting a fixed WiMAX network to the service-aware network is distributed between NKUA premises in Greece and VTT's CNL in Finland (Figure 7-6). The WiMAX network environment consists of Airspan MicroMAX BS [159] and Airspan ProST subscriber stations (SS) located on the VTT's CNL [160]. The testbed configuration is given in Table 7-2. The BS and SS operate in a laboratory environment with short distance direct line-of-sight conditions, resulting in stable and strong signal strength throughout the experiments. Regarding the fixed part of the testbed, the NKUA side has carried out experiments over the WiMAX testbed, remotely via the Internet, setting up IP routing and IP tunnelling (IPIP) from and to the WiMAX link. The Distributed Internet Traffic Generator (D-ITG) [161] generates traffic at both NKUA end machines. Traffic sender can concurrently generate multiple flows with user-defined parameters that can be analyzed from the receiver side, to extract traffic QoS features. Besides the VoIP traffic, TCP and UDP BT is additionally injected by D-ITG in order to realistically evaluate a VoIP traffic paradigm.

The NECM and NDCM introduced in Sections 4 and 7.1, as well as the algorithms proposed in Section 7.2 and 7.3 have been implemented using Java programming language. The BS control software (i.e., NECM) is implemented in order to dynamically collect WiMAX link information from the BS and to control QoS settings on the fly. The

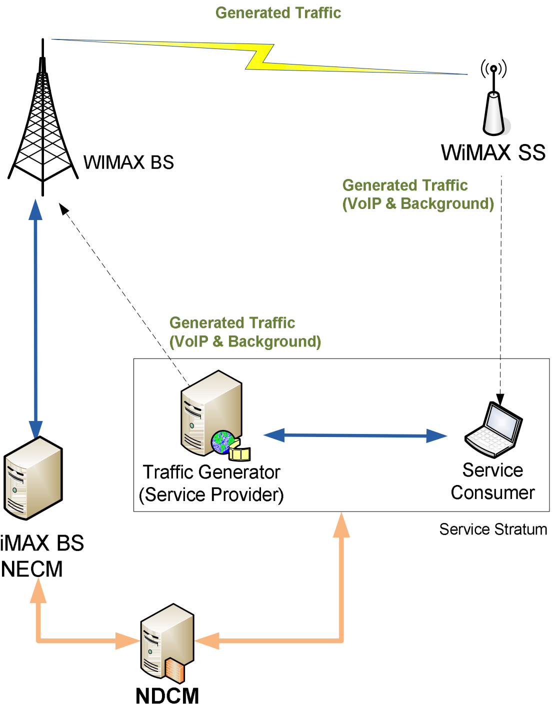


Figure 7-6: Testing Facilities

SNMP has been used for setting up a new configuration to the WIMAX BS and acquiring monitoring data. On the other side, the service level NECM interacts via SSH commands with the D-ITG and service consumers for monitoring and changing the codec of VoIP flows.

The change priority mechanism that has been deployed in the WiMAX BS is thoroughly described below. IEEE 802.16 standards specify various packet scheduling

Table 7-2: WiMAX Testbed Configuration

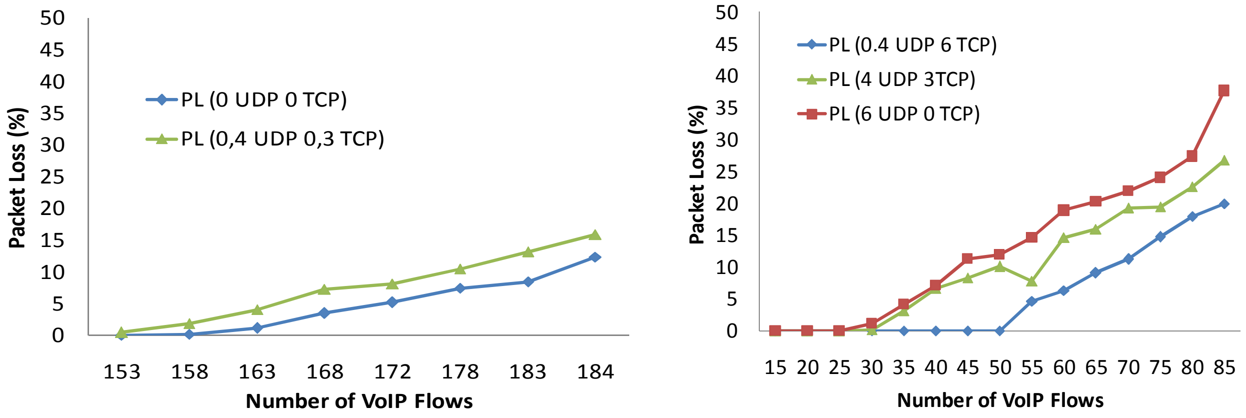
Testbed Parameter	Value
Base station	Airspan MicroMAX
Subscriber stations	Airspan ProST
PHY	IEEE 802.16d, 256 OFDM
Duplexing	Frequency Division Duplexing
Frequency	3.5 GHz
Bandwidth	3.5 MHz
Distance between BS and SS	10 m, direct line-of-sight
Modulation (UL and DL)	64 QAM (FEC: ¾)
CINR DL/UL	29 dB / 18 dB

schemes to ensure the required QoS for different traffic types. More specifically, IEEE 802.16d [162], in which the employed WiMAX testbed is based on, comprises of four different scheduling types, namely Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-real-time Polling Service (nrtPS), and Best Effort (BE). In the employed BS, the aforementioned scheduling types are supported only in the uplink through a request/grant scheduling. In the downlink, the BS supports a scheduling scheme, where several traffic flows can only be treated with different priorities, without assuring any delay or bandwidth requirements. This downlink scheduling type based on Weighted Fair Queuing (WFQ) is capitalized on our experiments. In WFQ, higher priority packets take priority over lower priority, but the fairness is achieved by servicing each queue fairly in terms of bit count. For instance, if the weightings of all queues are the same, the scheduling algorithm allows all queues to transmit the same amount of data in total independently of the packet sizes. Higher weighting provides a larger amount of bandwidth. In our case study, we used weightings of 80 and 50 for high and normal priority queues, respectively. According to WFQ equation $R\omega_1 / (\omega_1 + \omega_2)$, where R indicates the total bitrate and ω_i the weighting of queue *i*. The higher priority queue gets 62% of the total bandwidth. Packets in the individual queues are handled with the BE scheduling in the first-in first-out manner.

During the default scheduling, all traffic is treated equally by the packet classifier and placed to the same normal priority transmission queue, where the BE scheduling scheme picks packet for transmission. The BS controller can be prompted to configure the BS in order to handle particular traffic flows with higher priority. In this case the BS is dynamically configured to have two transmission queues of different priorities (i.e., normal and high priority). The packet classifier decides about the appropriate priority queue based on the packet's source and/or destination MAC address, IP address, or port number. The latter is used in our experiments for classifying the VoIP traffic flows.

7.5 Performance Evaluation

Various experiments have been performed by combining different number of VoIP flows and BT levels. The PL_{thr} is set to 1%, which indicates the acceptable maximum threshold for an efficient VoIP experience, as it is recommended by the WiMAX Forum [163]. Moreover, for the available codecs the defined weights of equation (2) are set as follows: W_1=1 for G.711.1, W_2=0.8 for G.711.2, W_3=0.6 for G.729.2, W_4=0.4 for G.729.3, and W_5=0.2 for G.723.1.



(a) **(b)**

Figure 7-7: Packet loss (a) Case 1, Case 2 of background traffic (b) Case 3, Case 4 and Case 5 of background traffic

The following cases for BT level in the network have been considered in the conducted experiments:

- Case 1: 0.0 Mbps of UDP data and 0.0 Mbps of TCP data
- Case 2: 0.4 Mbps of UDP data and 0.3 Mbps of TCP data
- Case 3: 0.4 Mbps of UDP data and 6.0 Mbps of TCP data
- Case 4: 4.0 Mbps of UDP data and 3.0 Mbps of TCP data
- Case 5: 6.0 Mbps of UDP data and 0.0 Mbps of TCP data

Figure 7-7-a and Figure 7-7-b depict the detected PL for the different number of flows and cases of the BT level. The priority of the VoIP flows is set as normal, while the codec type is equally allocated among the available VoIP flows (e.g., 150 VoIP flows consist of 30 G.711.1, 30 G.711.2, 30 G729.2, 30 G729.3, and 30 G.723.1 flows). As expected, the results show that the PL increases proportionally to the number of VoIP flows, while the higher is the BT, the less VoIP flows are needed to increase the PL above PL_{thr}.

7.5.1 Illustrative Examples for Adaptation Method Selection

Firstly, we study the selection of the adaptation action phase of the decision making algorithm, focusing on three of the conducted experiments. The maximum capacity of the WiMAX network is 9 Mbps (C_{WiMAX}) and PL_{max} is set to 6%. The introduced BT that traverses the network is 6.4 Mbps. In the first experiment the VoIP traffic is 1.584 Mbps and the detected PL (PL_g) is 6%, having marginal case (Table 7-3). Applying equation (7-6), it is estimated that the expected PL value (PL_g) could be reduced to 4.62%, using the change priority action. However, this is not considered as an efficient action since $PL_g \succ PL_{thr}$. For this reason, the modification of VoIP flows' codec is then selected

Table 7-3: Packet loss reduction after Change Priority Adaptation

VoIP Flows	$T_{VoIP\ (Mbps)}$	I	PL_d % (before adaptation)	PL_g (%)	PL_d % (after adaptation)
60	1.584	0.176	6.0	4.62	4.30
50	1.320	0.146	2.0	0.70	0.92
55	1.452	0.161	2.32	0.98	1.11

and the initial allocation of 12 flows per codec (i.e., totally 60 VoIP flows) is adapted as follows: 12 flows for G.711.1, 0 flows for G.711.2, 0 flows for G.729.2, 12 flows for G.729.3, and 36 flows for G.723.1 codec. Applying this adaption, the PL is reduced to 0.22%.

In the second case, less VoIP traffic has been injected in the network (1.32 Mbps), while the rest parameters remain the same (Table 7-3). PL_d is 2% and we estimate, via equation (6) that by applying the change priority action, the PL would be reduced below the desirable threshold (PL_g=0.70%$\prec PL_{thr}$). Hence, the next step includes the selection of the appropriate category as regards the number of the VoIP flows that should move from normal to high priority at the WiMAX BS queue. Solving equation (7-7) the following estimations for PL improvement are calculated: a) $t_k=1$: *PL =0.70%*, b) $t_k=1/2$: *PL =1.35%*, c) $t_k=1/3$: *PL =1.56%*. Thus, all VoIP flows (t_k=1) should be moved from normal to high priority, in order to reduce PL below 1%. After this adaptation the PL is reduced to 0.92%.

In the third experiment the injected VoIP traffic is 1.452 Mbps (Table 7-3). The decision making scheme selects the change priority action of all VoIP for addressing the 2.32% PL. However, after the change of VoIP flows' priority the detected PL is 1.11% and remains above PL_{thr}. In the next cycle of the NDCM control loop, the change of the codec of VoIP flows is selected. The initial allocation of 11 flows per available codec is adapted as follows: 11 flows for G.711.1, 0 flows for G.711.2, 9 flows for G.729.2, 2 flows for G.729.3, and 33 flows for G.723.1 codec. The PL reaches the level of 0.02%.

From the above discussion we notice that the proposed scheme provides a good approximation of the appropriate adaptation. Apart from the third experiment, where a second iteration and adaptation is required, since the initially selected change priority adaptation did not reduce the PL below 1%. The execution of an adaptation action that is not effective increases the transitory period for the communication system reconfiguration. Additional changes will be required, consuming more computational and communicational resources. This is because the defined $PL_{\max} = 6\%$ is not optimally set for all cases.

7.5.2 Self-Tuning of Decision Making Thresholds

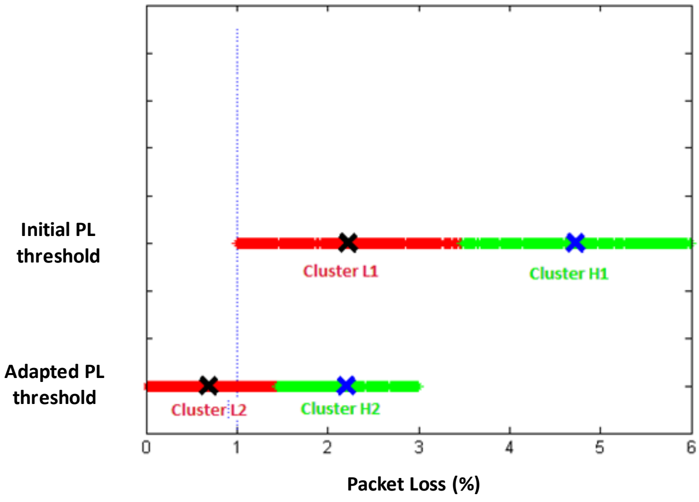
The PL instances that have triggered the change priority action are stored in the NDCM database. The unsupervised learning mechanism, introduced in Section 7.3, is applied to a batch of stored PL instances in order to dynamically tune the initially set value of $PL_{\max}$. Taking into account that PL_{thr}=1%, Figure 7-8-a visualizes the outcome of the application of k-Means algorithm on the stored PL instances. The calculated bounds and centroids of *L1*, *L2*, *H1*, and *H2* clusters are the following:

- L1 Centroid: 2.219 – L1 Bounds [1.002, 3.467]
- H1 Centroid: 4.721 – H1 Bounds [3.478, 5.995]
- L2 Centroid: 0.687 – L2 Bounds [0.004, 1.439]
- H2 Centroid: 2.203 – H2 Bounds [1.447, 2.999]

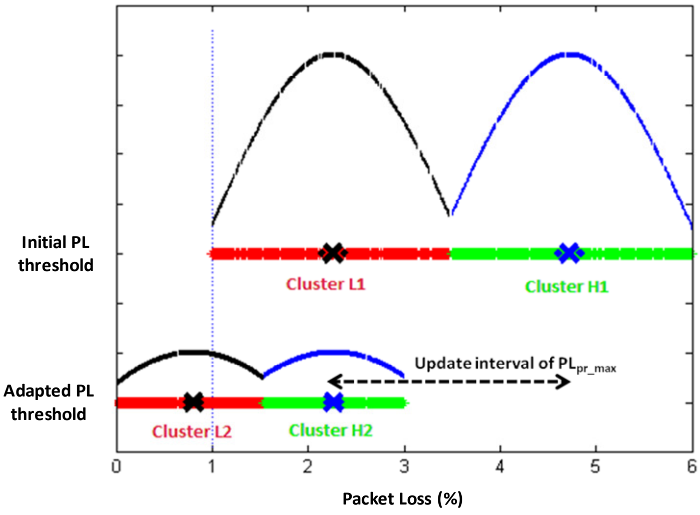
Based on Figure 7-8-b, we notice that the majority of *L1* cluster PL events have been successfully addressed, since the *L2* cluster is below PL_{thr}. On the other hand, *H1* PL events have been essentially decreased, but they have not reached PL_{thr}. Hence, PL_{max} needs to be decreased. The 90% confidence intervals of the cluster samples, using equation (16), are as follows:

- *L1:* [2.102, 2.337]
- *H1:* [4.604, 4.839]
- *L2:* [0.616, 0.758]
- *H2:* [2.132, 2.274]

Then, the updated PL_{max} is calculated. For the specific confidence intervals and sampling space, the updated threshold is $PL_{max} = 2.556\%$, as it is also deduced through Figure 7-8-b. The experiments presented in the previous section are repeated using the updated PL_{max} value (Table 7-4). The results show that PL_g is more accurately estimated, and specifically, in the third scenario the change of VoIP codec has been directly selected. The updated PL_{max} helps the cognitive manager to avoid redundant adaptations that extend and complicate the re-configuration period of the network.



(a)



(b)

Figure 7-8: Thresholds self-Tuning: (a) Clusters formed using k-Means Algorithm (b) PL Instance distribution and confidence intervals

VoIP Flows	$T_{VoIP\ (Mbps)}$	I	PL_d % (before adaptation)	PL_g (%)	PL_d % (after adaptation)
60	1.584	0.176	6.0	4.66	4.68
50	1.320	0.146	2.0	0.86	0.92
55	1.452	0.161	2.32	1.18	-

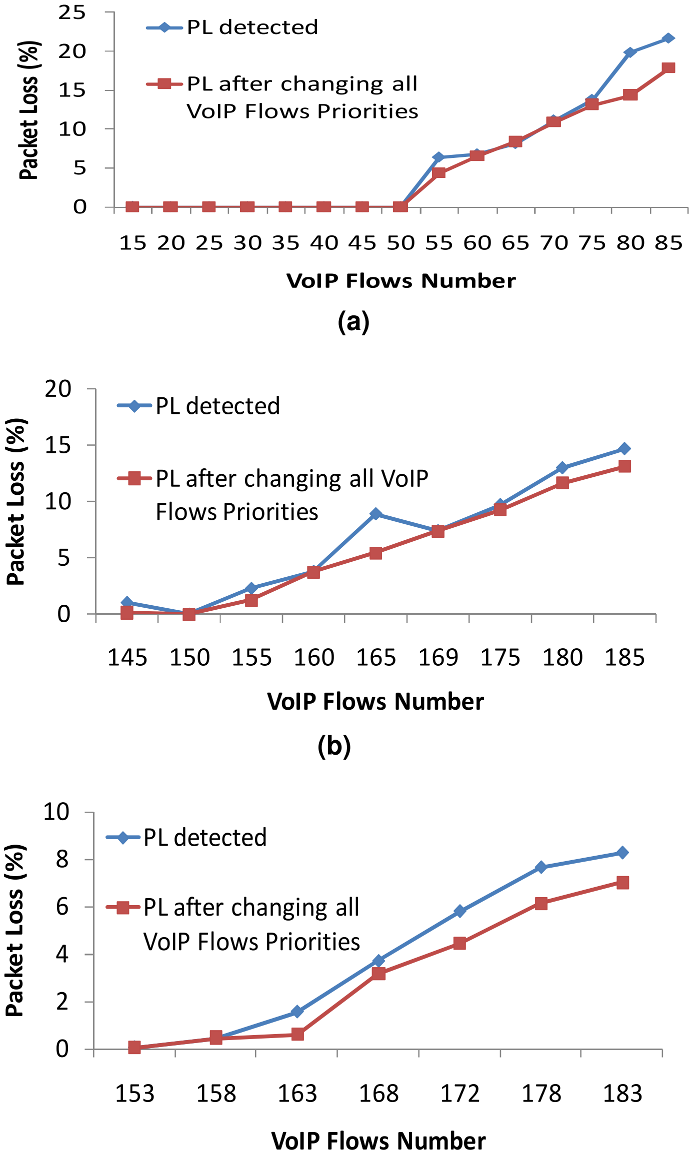
Table 7-4: Packet loss reduction after Change Priority Adaptation (After learning phase)

7.5.3 Packet Loss Reduction

The performance of the proposed configuration actions (change priority of VoIP flows at the WiMAX BS, change codec of VoIP flows) and the behavior of the system after each reconfiguration are further analysed below. Figure 7-9 shows that after the change of VoIP flows priority, the PL_{thr} is reached mainly for low PL events. As expected, the PL improvement is proportional to the number of the adapted flows (i.e., selected category:

all, 1/2, 1/3). Moreover, the effectiveness of a change priority action is reduced for a high BT level (e.g., Figure 7-9-b, Figure 7-9-c), without even reaching PL_{thr} in many cases. Hence, this data plane re-configuration is more appropriate for addressing low PL problems and it is important that it does not affect VoIP service continuity and the end user service experience.

As it analysed below, the modification of the codec of VoIP flows, which is service plane action, is more effective comparing to the change priority action, regardless the PL level. Experiments for the heuristic (Section 7.2.3.1) and the history-based (Section 7.2.3.2) schemes have been conducted using all the five cases of BT. In each experiment the total number of flows is initially equally allocated per codec type.



(a)

(b)

(c)

Figure 7-9: Change all VoIP flows priority(from normal to high) a) 0 Mbps UDP & 0 Mbps TCP BT, (b) 0.4 Mbps UDP & 0.3 Mbps TCP BT, (c) 0.4 Mbps UDP & 6 Mbps TCP BT

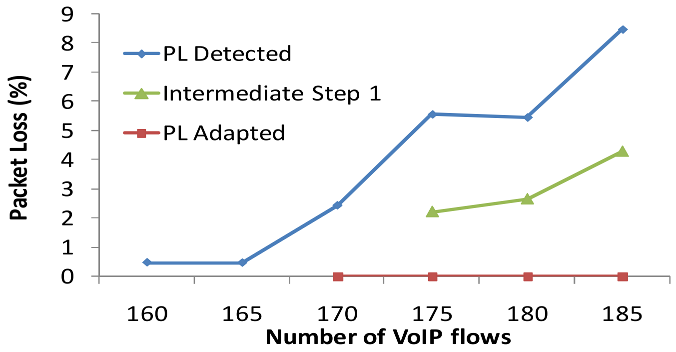
Table 7-5: Before Heuristic Codec Change (4 Mbps UDP, 3 Mbps TCP BT)

Initial VoIP Codec						PL & Service Cost	
Flows	**G711.1**	**G711.2**	**G729.2**	**G729.3**	**G723.1**	**PL(%)**	**SC**
35	7	7	7	7	7	3.56	21.00
45	9	9	9	9	9	9.05	27.00
60	12	12	12	12	12	19.42	36.00
85	17	17	17	17	17	26.42	51.00

Table 7-6: After Heuristic Codec Change (4 Mbps UDP, 3 Mbps TCP BT)

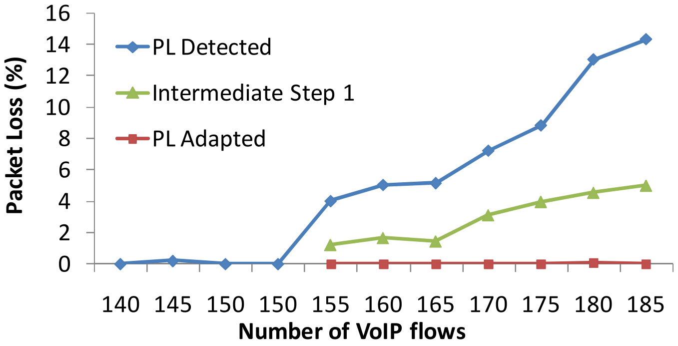
Adapted VoIP Codec						Adapted PL & Service Cost		
Flows	**G711.1**	**G711.2**	**G729.2**	**G729.3**	**G723.1**	**PL(%)**	**SC**	**Steps**
35	6	1	0	0	28	0.00	12.40	2
45	1	8	0	0	36	0.00	14.60	2
60	0	0	5	0	55	0.00	14.00	3
85	0	0	0	0	85	5.78	17.00	3

Figure 7-10 shows that the heuristic-based scheme provides very good adaptation results for relatively low PL and BT levels. However, as the detected PL and the injected BT volume are increasing, more iterations are required to reach PL_{thr}. For example, in a low level BT scenario, PL events that exceed 4% could be successfully addressed via the heuristic-based scheme using one additional iteration ('Intermediate Step 1' of Figure 7-10-b). For high BT conditions, as Figure 7-10-d shows, more steps are required to reduce the PL (four at maximum). In addition, Table 7-5 and Table 7-6 present the initial and the final status for some of the experiments of Figure 7-10-e, using the heuristic change codec scheme. The tables provide information for the allocation VoIP flows' codecs, the detected PL levels and the service cost, before and after change codec. A bad estimation of network conditions or the absence of knowledge about previous events increases the number of iterations that are needed to converge towards PL_{thr}. The heuristic scheme passes through a set of intermediate states, while moving from a given state to the target state. This phenomenon, namely bouncing effect, increases the transitory period of a communication system. The latter affects negatively efficient service delivery and increases the management overhead (i.e., computational and communication resources).



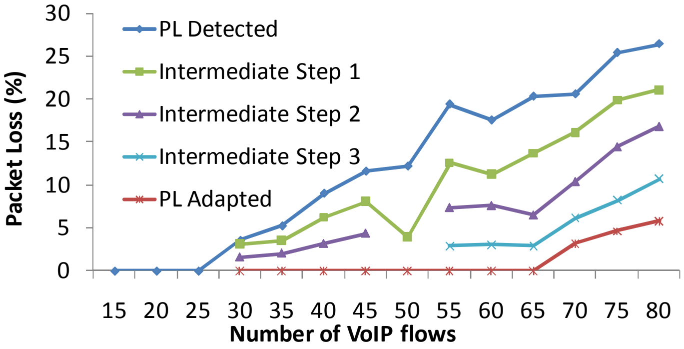
Packet Loss (%)
9
8
7
6
5
4
3
2
1
0
PL Detected
Intermediate Step 1
PL Adapted
160
165
170
175
180
185
Number of VoIP flows

(a)



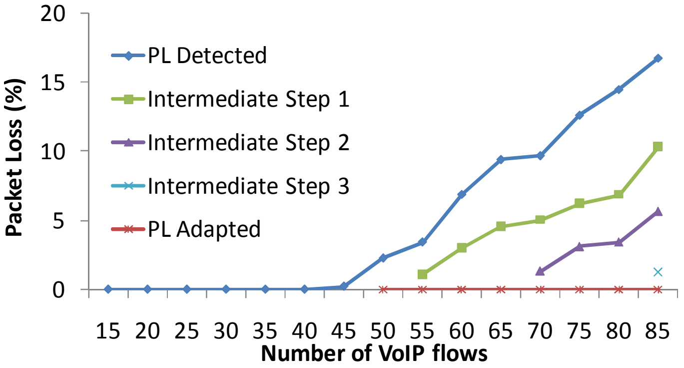
Packet Loss (%)
16
14
12
10
8
6
4
2
0
PL Detected
Intermediate Step 1
PL Adapted
140
145
150
150
155
160
165
170
175
180
185
Number of VoIP flows

(b)



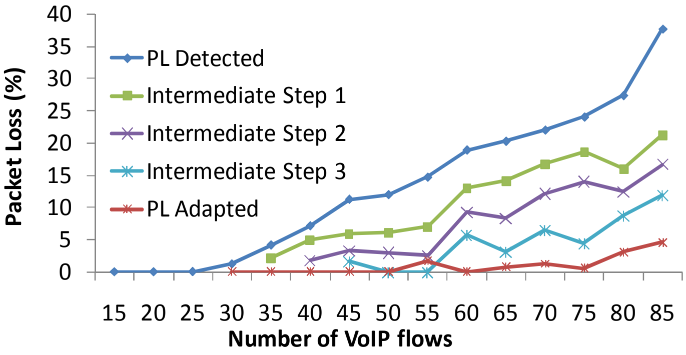
Packet Loss (%)
30
25
20
15
10
5
0
PL Detected
Intermediate Step 1
Intermediate Step 2
Intermediate Step 3
PL Adapted
15
20
25
30
35
40
45
50
55
60
65
70
75
80
Number of VoIP flows

(c)



Packet Loss (%)
20
15
10
5
0
PL Detected
Intermediate Step 1
Intermediate Step 2
Intermediate Step 3
PL Adapted
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
Number of VoIP flows

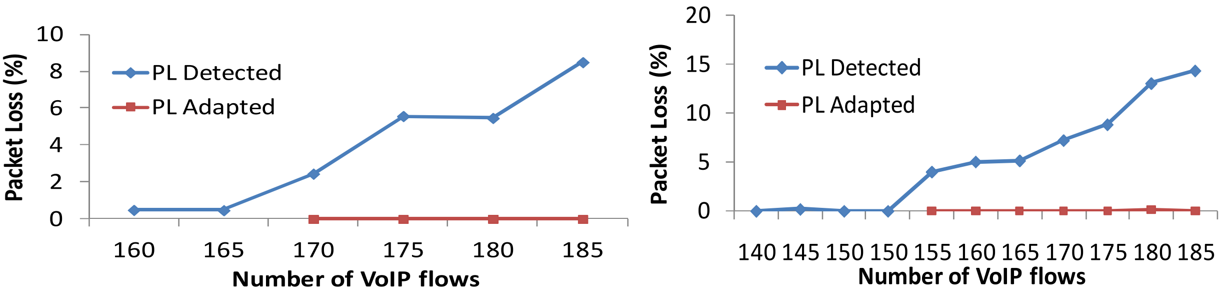
(d)



(e)

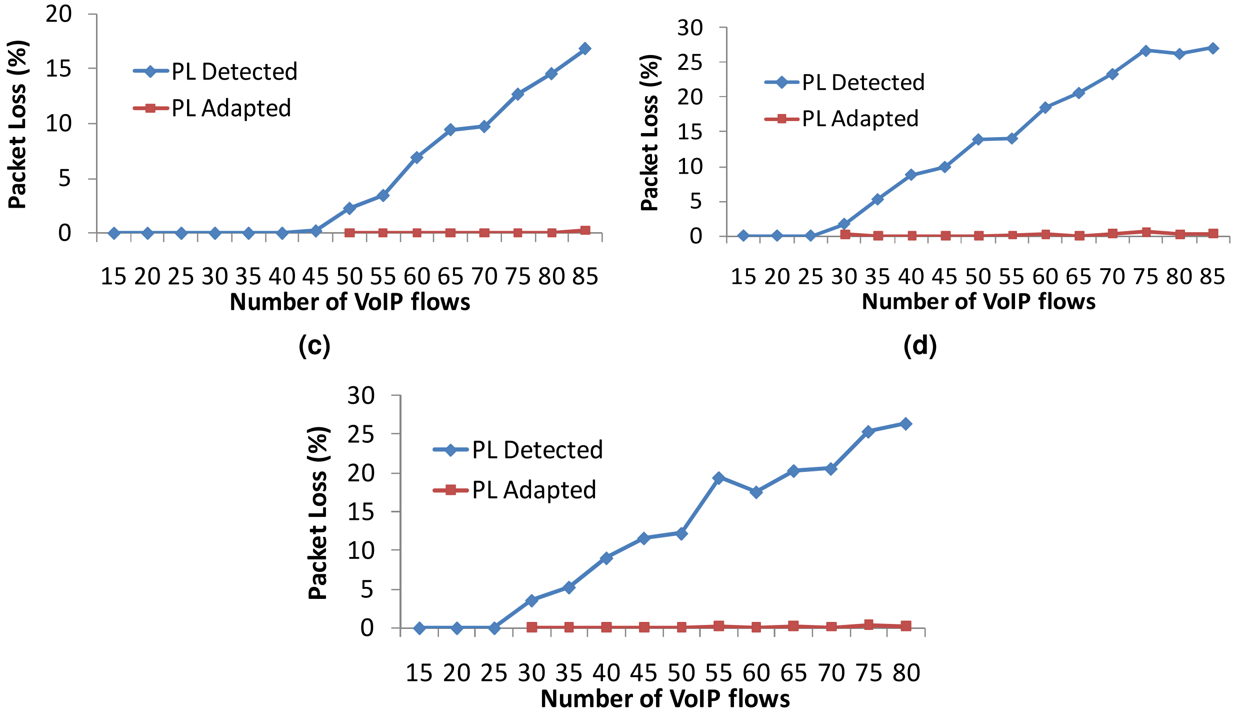
Figure 7-10: Heuristic Change of VoIP flows Codec (a) 0 Mbps UDP and 0 Mbps TCP BT, (b) 0.4 Mbps UDP and 0.3 Mbps TCP BT, (c) 0.4 Mbps UDP and 6 Mbps TCP BT, (d) 6 Mbps UDP and 0 Mbps TCP BT, (e) 4 Mbps UDP and 3 Mbps TCP BT

After the validation of the performance of the heuristic algorithm and having conducted numerous experiments for different BT levels and number of VoIP flows, the database of the NDCM contains a significant amount of information regarding the monitored PL level, the number of VoIP flows and the volume of the BT that traverses the network. This information is used by the history-based algorithm for the change of the codec of VoIP flows, as discussed in Section 7.2.3.2. The history of previous events and decided adaptations improves the performance and the convergence time of change codec re-configuration. The results of the conducted experiments are presented in Figure 7-11. We have used the same number of VoIP flows and BT traffic as in the heuristic approach. The most important benefit of this method is that no intermediate steps (i.e., iterations) are required to reach the desirable PL level (less than 1%). The history of previous events that is stored in the NDCM database (i.e., PL and Service Cost values for the specific BT level), in conjunction with the linear interpolant produce a concrete decision about SC_g that leads to a more efficient PL reduction. Figure 7-11 shows that this method performs very well for every combination of VoIP flows and BT. Consequently the transitory period of the system is minimized due to the absence of several iterations; facilitating the smoother adaptation of the communication network to the decided configuration.



(a)

(b)



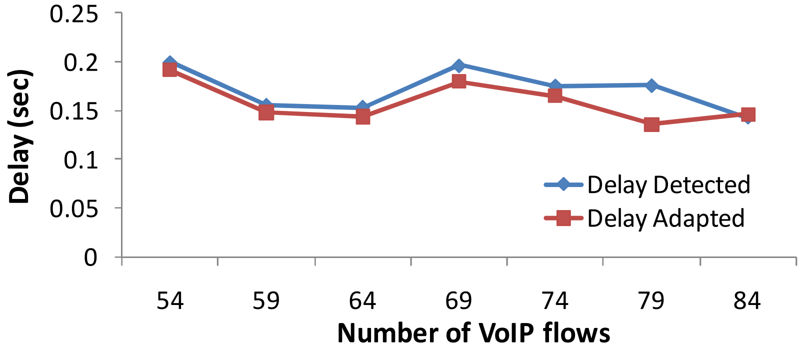
(c) (d)

(e)

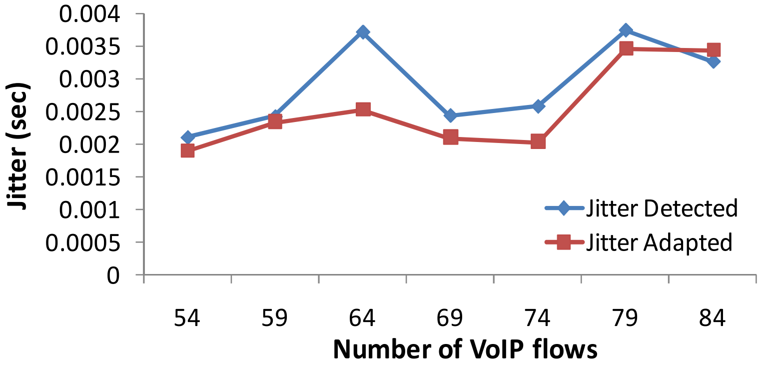
Figure 7-11: History-based Change of VoIP flows Codec (a) 0 Mbps UDP and 0 Mbps TCP BT, (b) 0.4 Mbps UDP and 0.3 Mbps TCP BT, (c) 0.4 Mbps UDP and 6 Mbps TCP BT, (d) 6 Mbps UDP and 0 Mbps TCP BT, (e) 4 Mbps UDP and 3 Mbps TCP BT

7.5.4 Delay, Jitter and R-Score

Apart from PL reduction, which is the main triggering parameter of the proposed solution, we discuss below the effect of the proposed adaptations on the average delay and jitter of the VoIP flows. Figure 7-12 and Figure 7-13 present the impact of a change priority and change codec action, respectively, on both performance metrics for a different number of VoIP flows. The BT consists of 0.4 Mbps of UDP data packets and 6 Mbps of TCP data packets. The adapted delay and jitter levels are reduced regardless of the type of the configuration action (Figure 7-12, Figure 7-13). However, similar to the PL case, the change of VoIP flows codec is also more effective for the reduction of delay and jitter, especially for a small number of VoIP flows (Figure 7-13).

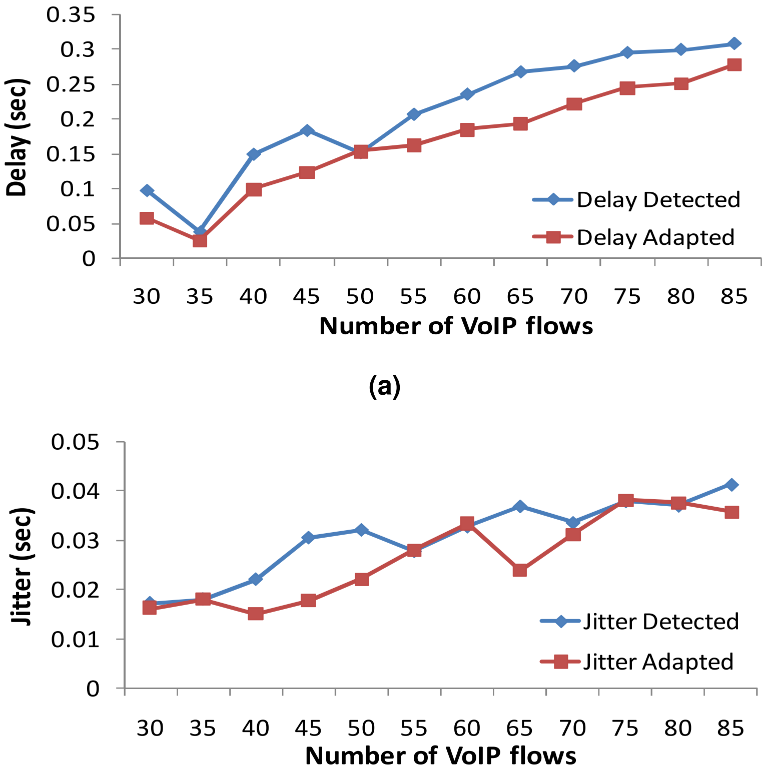


(a)



(b)

Figure 7-12: Change priority of all VoIP flows for 0.4 Mbps UDP and 6 Mbps TCP BT (a) Delay (b) Jitter



(b)

Figure 7-13: Change codec of VoIP flows for 0.4 Mbps UDP and 6 Mbps TCP BT (a) Delay (b) Jitter

The ITU E-Model [164] is an objective method to evaluate quality in VoIP systems, which we have also used. E-Model resulting score is the transmission rating R-factor which is calculated based on the combinational processing of a set of parameters, such as delay and packet loss [165], [166]. R-factor for VoIP service typically ranges from 0 (poor quality) to 100 (excellent quality); values below 60 are considered unacceptable. Using the measured delay and packet loss values, Figure 7-14shows the R-factor change, for a low PL case where the change priority action is enabled (Figure 7-9, Figure 7-12), as well as for high PL events where the history-based scheme for codec change is selected (Figure 7-11, Figure 7-13). In the specific experiments the BT contains 0.4Mbps

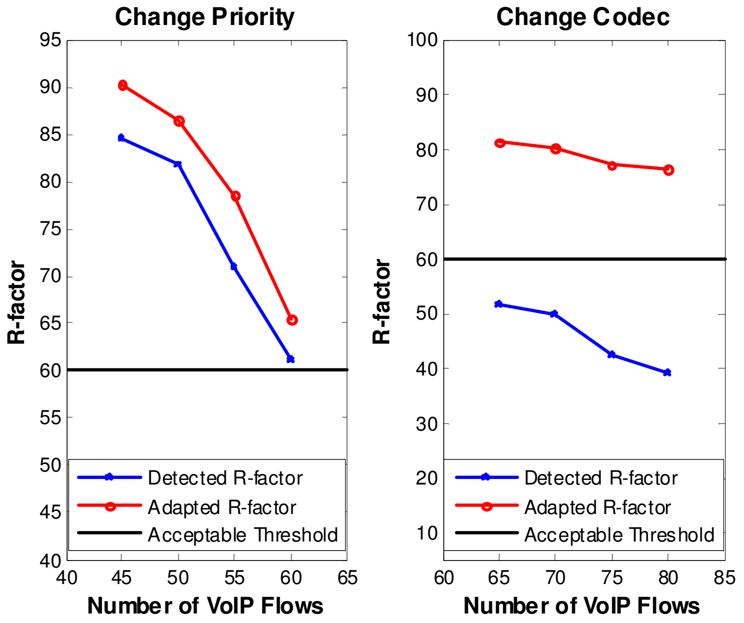


Figure 7-14: R-Factor after Change Priority and Change Codec adaptation

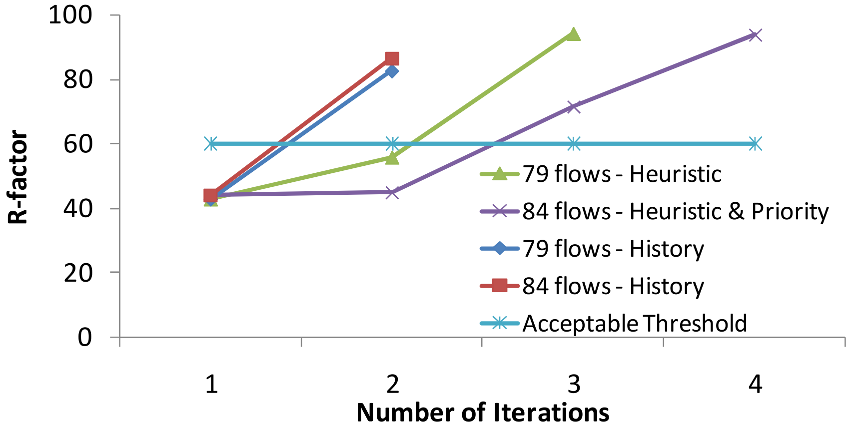


Figure 7-15: R-factor vs. Number of Iterations

of UDP and 6Mbps of TCP data packets and the decision making scheme is correctly tuned ($PL_{max}=2.556\%$). A change priority action increases the R-score by 5.7%-10.5%, while a change codec action improves the R-score by 57%-95%. In both cases the resulting R-score is higher than 60.

Figure 7-15 presents the change of R-factor values for two of the conducted experiments (79 VoIP flows, 84 VoIP flows) using both the heuristic and the history-based schemes. The BT is the same as above. Taking into account the measured PL and delay, we calculate the R-Factor, before and after the re-configuration action as well as for each intermediate step. In the heuristic scheme, the transitory period (i.e., in terms of iterations) is longer, since more steps are required for reaching PL_{thr}. The outcome of the heuristic scheme is better but the duration of the re-configuration period is two and three times longer, for 79 and 84 VoIP flows, respectively.

The conducted experiments, using FIRE Panlab and CORE facilities, prove the feasibility and the strengths of our work. Both network and service side adaptation actions improve the detected packet loss level. However, the change priority of VoIP flows at WIMAX BS cannot reach PL_{thr} for high PL events; although service continuity is always satisfied. On the other hand, the VoIP codec modification is more drastic. The utilisation of the history of previous adaptations reduces the transitory period and the iterations that a heuristic scheme requires for QoS improvement. The results of the learning phase show that the accuracy of the decision making scheme is improved, avoiding adaptations that are not effective. Finally, we have showed that other QoS metrics such as delay, jitter and R-factor are also improved by the proposed solution.

CHAPTER 8

CONCLUSIONS

Efficient usage of network resources and energy saving are important challenges for next generation wireless networks. They both reduce operational costs and contribute to the environmental protection. Even a small improvement in the energy consumption of a network node (e.g., UEs, AP) could become significant if it is scaled to hundreds or thousands of nodes. Energy saving in a wireless environment is a complex task, because of the interactions among the various types of nodes that constitute a network area, and within the different components that exist. Self-Organization enables a network to handle complexity, without any external or central dedicated control entity, and the spatio-temporal dynamics of a dense urban environment.

Initially, the functional and software architecture of the cognitive framework for self-organizing networks has been presented and its deployment in a realistic network environment is discussed. Then SYSTAS algorithm is proposed for the distributed discovery and establishment of clusters among network nodes that exist in a SON. The adoption of a cluster-based approach facilitates the cooperation and the coordination of a group of network nodes for identifying and solving local network management problems. The proposed algorithm has been evaluated using various graphs of network topologies (sparse, dense), leading to efficient discovery of clusters and calculated modularity metric.

We have proposed a cluster-based scheme for energy saving and CCO management tasks in a self-organized wireless LAN. The head of the cluster collects periodic measurements from APs, and calculating CCO and energy-related metrics selects the appropriate configuration action type for energy saving and resource management. A rule-based scheme has been presented for the selection of the appropriate adaptation in order to reach a more energy efficient state for the whole network area; the proposed scheme takes into account the utilization of network resources as well as the change in energy consumption after each adaptation and between monitoring periods. The proposed algorithms for CCO and energy saving configuration actions are: a) AP dynamic de/re-activation, b) UE multi-hop relays formation and c) AP frequency channel re-selection. Simulations have performed by using OPNET simulation environment, where the above mentioned schemes and algorithms have been developed, in order to analyze the CCO and energy consumption effect, at device and component level, as well as their interactions.

In this dissertation, we have focused on the communication component (Rx, Tx) of both APs and UEs in a dense WLAN network. The decomposition of energy consumption into device and component levels facilitates the analysis and the identification of their dependencies. From the obtained simulation results it is evident that coverage and capacity optimization is a tool for achieving energy efficiency. However, the extend of energy reduction as well as the impact on other system components (Rx, Tx) or nodes depends on the specific network topology, the network configuration features and traffic conditions. The simulation results based on the topology configuration presented above, show that an AP deactivation action, reduces cluster level EC_{AP}, mainly due to the reduced energy spent for packets reception (13% decrease) and processing. However, it increases the energy consumption of UEs for the reception and the transmission phase ($EC_{UE}^{Tx}, EC_{UE}^{Rx}$) under specific topology and traffic conditions (e.g., high overlap of frequency channels, UL/DL ratio). On the other hand, the consumed energy of UEs for the reception and the transmission phases increases after APs deactivation. The change of the selected channels in the cluster leads to the reduction of the energy that UEs consume for the sensing of data packets that come from neighboring cells. APs gain also benefit from this adaptation. EC_{UE}^{Rx} and EC_{AP}^{Rx} measured in nJ/bit are improved by 35.7% and 40.5%, correspondingly. Furthermore, the handover of a UE to a more distant AP leads to an increase of the energy used for transmissions (EC_{UE}^{Tx}), especially for a UE with high UL traffic. In this case, the formation of UEs multi-hop relays improves the energy consumption in the data packet transmission phase by 20%.

In the conducted experiments that described above, the degradation of a performance metric, after the enforcement of a re-configuration is addressed by using an additional optimization action. However, in the case that two or more configurations actions are triggered concurrently it is important for the SON to identify and resolve conflicts on metrics or parameters, so as to assure the stability of the communication network. A novel scheme that consists of three phases is proposed for the coordination of a SON: a) a time series-based approach is used to avoid checking a configuration action that is triggered by a performance metric, which value appears temporary fluctuations, b) Identify and resolve conflicts on a triggered configuration action, according to the severity of the problem and the priority of the respective performance metric, c) Identify and resolve conflicts on performance metrics using a cost-benefit analysis method. The proposed solution has been implemented and applied in a wireless network environment, where CCO, energy saving and interference control use cases are handled by the proposed SON Coordination solution. Performance improvement through conflicts and dependencies resolution has been evaluated using OPNET simulation tool.

Moreover, in this dissertation we have addressed the issue of VoIP QoS assurance in wireless network environment using cognitive self-management techniques. The solution builds on the cooperation between NMS and service stratum, utilizing cognition and self-management enablers. We have proposed an algorithmic framework that is deployed on the agents of the introduced self-management architecture (NECM, NDCM), extending their cognitive capabilities. The most appropriate adaptation is selected, by assessing the PL level, the cost of VoIP service and other traffic conditions. The modification of the VoIP flows priority at the network side (i.e., WIMAX BS) and the change of selected VoIP codec (i.e., service side) are the studied adaptations that provoke network and service stratums interaction. The history of previous events is utilized by the cognitive network manager, wherever the former is available. Otherwise heuristic schemes are activated. The evolution of the decision making scheme, using a feedback-based technique is a key capability of our solution.

The conducted experiments, using FIRE Panlab and CORE facilities, prove the feasibility and the strengths of our work. Both network and service side adaptation actions improve the detected packet loss level. However, the change priority of VoIP flows at WIMAX BS cannot reach the target PL threshold (<1%) for high PL events; although service continuity is always satisfied. On the other hand, the VoIP codec modification is more drastic. The utilization of the history of previous adaptations reduces the transitory period and the iterations that a heuristic scheme requires for QoS improvement. The results of the learning phase show that the accuracy of the decision making scheme is improved, avoiding adaptations that are not effective. Finally, we have showed that other QoS metrics such as delay, jitter and R-factor are also improved by the proposed solution.

The following list presents some open research issues for future work:

- Extend SYSTAS clustering algorithm, inserting weights for each edge, according to specific criteria e.g., distance among APs.
- Extend the work presented in this dissertation for a mobile network environment.
- Enrich the list of configuration actions (e.g., AP power control) and performance metrics (e.g., EC_{AP}^{Tx}) for energy saving in a self-organized wireless network.
- Extend the list of configuration actions and performance metrics for the

investigation of SON coordination in a wireless network environment.

- Study and experimentation on a probabilistic scheme for the resolution of dependencies among different optimization problems in SONs.

ABBREVIATIONS – ACRONYMS

3GPP	3rd Generation Partnership Program
ACE	Algorithm for Cluster Establishment
ACT-R	Adaptive Character of Thought—Rational
AM	Autonomic Manager
ANA	Autonomic Network Architecture
AP	Access Point
API	Application Programming Interface
BE	Best Effort
BER	Bit Error Rate
BFS	Breadth First Search
BS	Base Station
CA	Configuration Action
CAPEX	Capital Expenditures
CC	Clustering Coefficient
CCO	Coverage and Capacity Optimization
CDR	Current Data Rate
CEC	Current Energy Consumption
CHOP	Configure Heal Optimize Protect
CNM	Cognitive Network Manager
COOP	Coverage Optimization Opportunity
CPU	Central Processing Unit
CS	community structure
CUR	Capacity Usage Ratio
DCA	Distributed Clustering Algorithm
DCSNM	Distributed Cognitive cycle for System & Network Management
DF	Dominance Factor

DL	Downlink
DMAC	Distributed Mobility Adaptive Clustering Algorithm
EC	Energy Consumption
EDR	Expected Data Rate
EEC	Expected Energy Consumption
eNB	Evolved Node B
FCAPS	Fault, Configuration, Accounting, Performance, and Security
FIRE	Future Internet Research and Experimentation
FL	Fuzzy Logic
GP	Graph partitioning
HCC	Hierarchical control clustering
HO	Handover Optimization
IC	Interference Cost
ICIC	Inter-Cell Interference Coordination
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INM	In-Network Management
ITU	International Telecommunication Union
LCA	Linked Cluster Algorithm
LCC	Least Cluster Change
LEACH	Low energy adaptive clustering hierarchy
LTE	Long Term Evolution
MAC	Medium Access Control
MANET	Mobile Ad-hoc Network
MIB	Management Information Base
NECM	Network Element Cognitive Manager
NDCM	Network Domain Cognitive Manager
NMS	Network Management System

OAM	Operations, Administration, and Maintenance
OF	Overlapping Factor
OPEX	Operational Expenditures
OSF	Operations Systems Functions
OSGi	Open Services Gateway initiative
OWL	Web Ontology Language
P2P	Peer-to-Peer
PEACH	Power-Efficient and Adaptive Clustering Hierarchy
PER	Packet Error Rate
PL	Packet Loss
PM	Performance Metrics
PRB	Physical Resource Blocks
QoS	Quality of Service
RACH	Random Access CHannel
RAT	Radio Access Technology
RF	Radio Frequency
RRM	Radio Resource Management
Rx	Received data
SAE	System Architecture Evolution
SASE	Self-Aware and Self- Effective
SCP	Service Cost and Packet Loss
SNMP	Simple Network Management Protocol
SON	Self-Organizing Network
SWRL	Semantic Web Rule Language
TCP	Trasmission Control Protocol
TMN	Telecommunications Management Network
TTL	Time-To-Live
Tx	Transmitted data

UDP	User Datagram Protocol
UE	User Equipment
UGS	Unsolicited Grant Service
UL	Uplink
VoIP	Voice over Internet Protocol
WFQ	Weighted Fair Queuing
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Network
WWW	World Wide Web

REFERENCES

[1] Cisco 2011a, “Global Mobile Data Traffic Forecast Update, 2010–2015”, February 2011.
[2] Nokia Siemens Networks 2011, “2020: Beyond 4G Radio Evolution for the Gigabit Experience”, February 2011.
[3] J. Hourcade, Y. Neuvo, R. Posch, et al., “Future Internet 2020: Visions of an Industry Expert Group”, European Commission Information Society and Media, May 2009.
[4] G. Fettweis and E. Zimmermann, “ICT Energy consumption – Trends and Challenges”, in Proc. IEEE WPMC, Lapland, Finland, Sept.2008.
[5] T. Chen, H. Zhang, Z. Zhao, X. Chen, “Towards green wireless access networks”, in Proc. CHINACOM, Beijin, China, Aug. 2010, pp.1-6.
[6] N. Agoulmine, ed., Autonomic network management principles: From concepts to applications, Elsevier, 2011.
[7] WWRF 2011, “Visions and research directions for the Wireless World”, October 2011.
[8] 3GPP TS 32.500, “Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements“, 2011.
[9] T. Jansen, M. Amirijoo, U. Turke, et al., “Embedding multiple self-organization functionalities in future radio access networks”, in Proc. IEEE GLOBECOM, Hawaii, Dec. 2009, pp.1-5.
[10] NGMN Alliance Requirement Specification, “NGMN Recommendation on SON and O&M Requirements”, edited by Frank Lehser, http://www.ngmn.org, Dec. 2008.
[11] 3GPP TR 32.826, “Telecommunication management; Study on Energy Savings Management (ESM)”, 2010
[12] T. Jansen, M. Amirijoo, U. Türke, et al.,Embedding multiple self-organisation functionalities in future radio access networks”, in Proc. IEEE VTC, Barcelona, Spain, Apr. 2009, pp.1-5.
[13] C. Courcoubetis and R. Weber, eds., Pricing Communication Networks: Economics, Technology and Modelling, Wiley, 2003.
[14] R.W. Thomas, “Cognitive Networks”, Ph.D. Dissertation, Virginia Polytechnic and State University, Blacksburg, VA, June 15, 2007.
[15] C. Prehofer, C. Bettstetter, “Self-organization in communication networks: principles and design paradigms”, IEEE Communication Magazine, Vol. 43, Issue: 7, pp. 78 – 85, 2005.
[16] Self-Organizing Networks, SON, tutorial, http://www.radio-electronics.com/info/cellulartelecomms/self-organising-networks-son/basics-tutorial.php
[17] ITU-T, “M.3010: Principles for a telecommunications management network”, 2000.
[18] RFC 3411, “An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks”, 2002.
[19] RFC 3418, “Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)”, 2002.
[20] R. Boutaba and A. Polyrakis, “Projecting advanced enterprise network and service management to active networks”, in IEEE Network Mag, vol. 16, no. 1,pp. 28-33, 2002.
[21] S. Calo and M. Sloman, “Guest Editorial: Policy-Based Management of Networks and Services”, J. Network and Systems Management, vol. 11, n. 3, pp. 249–252, 2003.
[22] A. Karmouch, “Mobile software agents for telecommunications”, IEEE Commun. Mag., vol. 36, no. 7, 1998, pp.24-25.
[23] A. M. Hadjiantonis and G. Pavlou, “Policy-based self-management of hybrid ad hoc networks for dynamic channel configuration,” in Proc. 11th IEEE/IFIP Network Operations and Management Symposium (NOMS). Salvador - Bahia - Brazil, April 2008, pp. 433-440.
[24] D. Cliff, “Biologically-Inspired Computing Approaches To Cognitive Systems: a partial tour of the literature”, Technical Report, http://www.hpl.hp.com/techreports/2003/HPL-2003-11.html, 2003.
[25] D. Vernon, G. Metta, G. Sandini, “A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents”, IEEE Transactions on Evolutionary Computation, vol.11, no.2, pp.151-180, April 2007.
[26] P. Langley, J. E. Laird, S. Rogers, “Cognitive architectures: Research issues and challenges”, (http://cll.stanford.edu/~langley/papers/final.arch.pdf), 2006.
[27] SOAR, http://sitemaker.umich.edu/soar/home
[28] ACT-R, http://act-r.psy.cmu.edu
[29] J. Weng, “On Developmental Mental Architectures”, Neurocomputing, vol. 70, no. 13-15, pp. 2303-2323, 2007.
[30] R.W. Thomas, L.A. DaSilva, A.B. MacKenzie, “Cognitive networks” in Proc. 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2005, pp.352–360.
[31] J. Mitola, III and G.Q. Maguire, Jr., "Cognitive radio: Making software radios more personal," IEEE Personal Communications, vol. 6, no. 4, pp. 13-18, 1999.

[32] Francisco J. Varela, Ed., “Principles of Biological Autonomy”, North Holland, Amsterdam, 1979.
[33] D.D. Clark, C. Partrige, J.C. Ramming, J.T. Wroclawski, “A knowledge plane for the Internet”, in Proc. SIGCOMM 2003, Karlsruhe, Germany, 2003, pp. 3-10.
[34] J. O. Kephart and D. M. Chess, “The vision of autonomic computing. Computer”, vol. 36, Issue: 1 pp. 41–52, 2003.
[35] R. Quitadamo and F. Zambonelli, “Autonomic communication services: a new challenge for software agents”. In Journal of Autonomous Agents and Multi-Agent Systems, vol. 17, no. 3, pp. 457-475, 2008.
[36] S. Dobson, S. Denazis, A. Fernández, et al., "A survey of autonomic communications", ACM Transactions on Autonomous and Adaptive Systems, vol 1, no. 2, pp. 223-259, 2006.
[37] J. K. Strassner, “Autonomic Systems and Networks: Theory and Practice”, In Proc. 10th IEEE/IFIP Network Operations and Management Symposium (NOMS), 2006, pp. 588-588.
[38] C. Fortuna and M. Mohorcic, “Trends in the development of communication networks: Cognitive networks”, In Computer Networks vol. 53, no., pp. 1354-1376, 2009.
[39] J.O. Kephart, “Research challenges of autonomic computing”, In Proc. 27th International Conference on Software Engineering, ICSE05, pp. 15-22, 2005.
[40] M. Parashar, and S. Hariri, “Autonomic Computing: An Overview”, UPP 2004, Mont Saint-Michel, France, Editors: J.-P. Banâtre et al. LNCS, Springer Verlag, vol. 3566, pp. 247-259, 2005.
[41] S. Hariri, B. Khargharia, H. Chen, et al., "The Autonomic Computing Paradigm", Kluwer Cluster Computing, Vol. 9, No. 1, pp. 5-17, 2006.
[42] A. Greenberg, G. Hjalmtysson, D.A. Maltz, et al., “A clean slate 4D approach to network control and management,” ACM SIGCOMM Computer Communication Review, vol. 35, no. 5, pp. 41-54, 2005.
[43] 4WARD, http://www.4ward-project.eu
[44] ANA, http://www.ana-project.org
[45] A. Galis, S. Denazis, A. Bassi, et al., “Management Architecture and Systems for Future Internet Networks”, in Towards the Future Internet – A European Research Perspective, G. Tselentis, J. Domingue, A. Galis, et al., Eds, Netherlands: IOS Press, pp. 95-104, 2009.
[46] Bionets, http://www.bionets.eu
[47] A. Manzalini, F. Zambonelli, “Towards Autonomic and Situation-Aware Communication Services: the CASCADAS Vision”. Proceedings of the IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications (DIS’06), 2006, pp. 383-388.
[48] H. Ballani and P. Francis, “CONMan: a step towards network manageability,” ACM SIGCOMM Computer Communication Review, vol. 37, no. 4, pp. 205-216, 2007.
[49] K.Nolte, A.Kaloxylos, K.Tsagkaris, et al, “The E3 architecture: Enabling future cellular networks with cognitive and self-x capabilities”, International Journal Of Network Management, vol. 21, no. 5, pp. 360–383, 2011.
[50] B. Jennings, S. Van der Meer, S. Balasubramaniam, et al., “Towards Autonomic Management of Communications Networks”, IEEE Commun. Mag. vol. 45, no. 10, pp. 112–121, 2007.
[51] R. Chaparadza, S. Papavassiliou, T. Kastrinogiannis, et al., “Creating a viable Evolution Path towards Self-Managing Future Internet via a Standardizable Reference Model for Autonomic Network Engineering”, in Towards the Future Internet - A European Research Perspective, G. Tselentis, J.Domingue A. Galis, A. Gavras, et al., Eds, Netherlands: IOS Press, 2009, pp. 136 - 147.
[52] Y. Yemini, A. Konstantinou, D. Florissi, “NESTOR: an architecture for network self-management and organization”, IEEE Journal on Selected Areas in Communications, vol.18, no.5, pp.758-766, 2000.
[53] A. Kousaridas, G. Nguengang, J. Boite, et al., “An experimental path towards Self-Management for Future Internet Environments”, in Towards the Future Internet - Emerging Trends from European Research, G. Tselentis, A. Galis, A. Gavras, et al., Eds, Netherlands: IOS Press, 2010, pp. 95 – 104.
[54] Socrates FP7 Project, http://www.fp7-socrates.org
[55] N. Samaan and A. Karmouch, "Towards Autonomic Network Management: an Analysis of Current and Future Research Directions", IEEE Trans. Commun. Surveys & Tutorials, vol. 11, no. 3, Aug. 2009, pp. 22-36.
[56] Z. Movahedi, M. Ayari, R. Langar, et al., “A Survey of Autonomic Network Architectures and Evaluation Criteria”, IEEE Trans. Commun. Surveys & Tutorials, vol. 14, no. 2, May 2012, pp. 464-490.
[57] S. Szott, M. Wódczak, R.Chaparadza, T. Meriem, K. Tsagkaris, A. Kousaridas et al., “Standardization of an Autonomicity-Enabled Mesh Architecture Framework, from ETSI-AFI Group perspective: Work in Progress”, In Proc. IEEE Globecom Workshop on Management of Emerging Networks and Services, 2012.
[58] M. Wadczak, Tayeb Ben Meriem, B. Radier, R. Chaparadza, K.Quinn, J. Kielthy, B. A. Lee, L. Ciavaglia, K. Tsagkaris, S. Szott, A. Zafeiropoulos, A. Liakopoulos, A. Kousaridas, M. Duault, “Standardizing a reference model and autonomic network architectures for the self-managing future internet” IEEE Network vol. 25, no. 6, pp. 50-56, 2011.

[59] D. Dudkowski, B. Tauhid, G. Nunzi, et al., "A prototype for in-network management in NaaS-enabled networks", IEEE/IFIP Int'l Symp. on Integrated Network Management, May. 2011, pp. 81-88.

[60] B. Biskupski, J. Dowling, J. Sacha, “Properties and mechanisms of self-organizing MANET and P2P systems”, ACM Trans. on Autonomous and Adaptive Systems (TAAS), vol. 2, no. 1, 2007.

[61] F. Dressler, "Self-Organization in Ad Hoc Networks: Overview and Classification," University of Erlangen, Dept. of Computer Science 7, Technical Report 02/06, March 2006.

[62] K. L. Mills, “A brief survey of self-organization in wireless sensor networks”, Wiley InterScience Wireless Communications and Mobile Computing, 2007, pp. 823-834.

[63] D. J. Baker and A. Ephremides, “The architectural organization of a mobile radio network via a distributed algorithm”, IEEE Trans. on Communications, vol.29, no.11, pp. 1694- 1701, 1981.

[64] T. Robertazzi, and P. Sarachik, "Self-organizing communication networks," IEEE Comm. Mag., vol.24, no.1, pp. 28-33, Jan 1986.

[65] L. Jiandong, L. Antao, “An adaptive cluster algorithm for a self-organizing communication network”, in Proc. IEEE GLOBECOM, 1988, pp. 1653–1656.

[66] A.G. Spilling, A.R. Nix,"Aspects of self-organisation in cellular networks," in Proc. IEEE PIMRC, 1998, pp. 682-686.

[67] Tonguz, Ozan K., S. Dixit, Sudhir, "Self-Organization in Cellular Wireless Networks via Fixed Relay Nodes", in Proc. IEEE GLOBECOM, 2006, pp.1-5.

[68] P. Jiang, J. Bigham, J. Wu, “Self-organizing relay stations in relay based cellular networks”, J. Elsevier Computer Communications, vol. 31, no. 13, pp. 2937-2945, 2008.

[69] K. Zimmerman, “An Autonomic Approach for Self-Organising Access Points,” Diploma Thesis, University of Ulm, Germany, 2004.

[70] L. Weidong, R.V. Prasad, A. Lo A, I, Niemegeers, “A Framework for the Self-organization of Personal Networks”, In proc. MobiQuitous, 2007, pp. 1-8.

[71] 3GPP TS 32.521 “Telecommunication management; Self-Organizing Networks (SON); Self-optimization and self-healing; Concepts and requirements”.

[72] M. Amirijoo, P. Frenger, F. Gunnarsson, et al., “On self-optimization of the random access procedure in 3G Long Term Evolution”, in Proc. IFIP/IEEE International Symposium on IM, 2009, pp. 177-184.

[73] NGMN Alliance, “Next Generation Mobile Networks Use Cases related to Self Organising Network, Overall Description,” v.2.02, December 2008.

[74] IEEE 802.16 Task Group m, http://wirelessman.org/tgm

[75] IEEE 802.11 Task Group s, http://grouper.ieee.org/groups/802/11/Reports/tgs_update.htm

[76] 3GPP TR 36.902, v9.3.1, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions”

[77] G. Anastasi, M.Conti, M. D. Francesco, A. Passarella, “Energy conservation in wireless sensor networks: A survey”, Elsevier J. Ad Hoc Networks, vol. 7 no. 3, pp. 537-568, May 2009.

[78] R. K. Sharma, C. L. Bash, C. D. Patel, et al., “Balance of Power: Dynamic Thermal Management for Internet Data Centers”, J. IEEE Internet Comput., vol. 9, no 1, pp. 42–49, Feb. 2005.

[79] R. Bolla, , R. Bruschi, F. Davoli, et al., “Energy Efficiency in the Future Internet: A Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructures”, IEEE Trans. Commun. Surveys & Tutorials, vol. 13, no. 2, pp. 223-244, Jun. 2011.

[80] C. Yan, Z. Shunqing, S. Xu, G.Y. Li, “Fundamental trade-offs on green wireless networks”, IEEE Commun. Mag, vol. 49, no. 6, pp. 30-37, Jun. 2011.

[81] L.M. Correia, D. Zeller, O. Blume, et al., “Challenges and enabling technologies for energy aware mobile radio networks”, IEEE Commun. Mag, vol. 48, no. 11, pp. 66-72, Nov. 2010.

[82] H. Congzheng, T. Harrold, S. Armour, et al., “Green radio: radio techniques to enable energy-efficient wireless networks”, IEEE Commun. Mag, vol. 49, no. 6, pp. 46-54, Jun. 2011.

[83] Energy Aware Radio and neTworking technologies (EARTH) FP7 EC funded project, https://www.ict-earth.eu

[84] G. Auer, V. Giannini, I. Godor, et al., “Cellular Energy Efficiency Evaluation Framework”, in Proc. IEEE VTC, Budapest, May 2011, pp.1-6.

[85] E. C. Strinati, A. D. Domenico, L. Herault, “Green Communications: an Emerging Challenge for Mobile Broadband Communication Networks”, J. of Green Engineering, vol. 1, no. 3, pp. 267-301, Mar. 2011.

[86] K. Samdanis, T. Taleb, D. Kutscher, et al, "Self Organized Network Management Functions for Energy Efficient Cellular Urban Infrastructures”, Springer J. MONET, vol. 17, no. 1, pp. 119-131, Feb. 2012.

[87] J. Lorincz, A. Capone, M. Bogarelli, “Energy savings in wireless access networks through optimized network management”, in Proc. IEEE ISWPC, Modena, Italy, May 2010, pp.449-454.

[88] F. Richter, A. J. Fehske, G.P. Fettweis, “Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks”, in Proc. IEEE VTC, Alaska, USA, Sept. 2009, pp.1-5.

[89] D. Cao, S. Zhou, C. Zhang, et al., “Energy Saving Performance Comparison of Coordinated Multi-Point Transmission and Wireless Relaying”, in Proc. IEEE GLOBECOM, Florida, USA, Dec. 2010, pp.1-5.

[90] A. P. Jardosh, K. Papagiannaki, E. M. Belding, et al., “Green WLANs: On-Demand WLAN Infrastructures”, Springer J. MONET, vol. 14, no 6, pp. 798–814, Dec. 2009.

[91] C. Comaniciu, N.B. Mandayam, H.V. Poor, “Radio Resource Management for Green Wireless Networks”, in Proc. IEEE VTC, Alaska, USA, Apr. 2009, pp.1-5.

[92] M. Gruber, S. Borst, E. Kuehn, “Stable Interaction of Self-Optimization Processes in Wireless Networks”, in Proc. IEEE ICC, Kyoto, Japan, Jun. 2011, pp.1-6.

[93] L.C. Schmelz, M. Amirijoo, A. Eisenblätter, et al., “A Coordination Framework for Self-Organisation in LTE Networks”, in Proc. IFIP/IEEE IM, 2011, pp. 193-200.

[94] I. Koutsopoulos and L. Tassiulas, “Joint optimal access point selection and channel assignment in wireless networks”, IEEE/ACM Trans. Netw., vol. 15, no. 3, pp. 521-532, 2007.

[95] E.C. Lupu and M.S. Sloman, “Conflicts in Policy-Based Distributed Systems Management” IEEE Trans. on Software Engineering, vol. 25, no. 6, pp. 852-869, 1999.

[96] M. Charalambides, G. Pavlou, P. Flegkas, et al., “Policy conflict analysis for diffserv quality of service management”, IEEE Trans. on Network and Service Management (TNSM), vol. 6, no. 1, pp. 15-30, 2009.

[97] S. Davy, B. Jennings and J. Strassner, “The policy continuum – Policy authoring and conflict analysis', in ACM Computer Communications, vol. 31, no. 13, pp. 2981-2995, 2008.

[98] Univerself FP7 Project, www.univerself-project.eu

[99] FP7 Univerself Deliverable D3.4, “Cooperation Strategies and Incentives”, December 2011.

[100] FP7 Socrates Deliverable D2.4 “Framework for the Development of self-Organising Networks”, 2008.

[101] X. Gelabert, B. Sayrac, and S.B. Jemaa, "A performance evaluation framework for control loop interaction in Self Organizing Networks”, in Proc. IEEE PIMRC, Sept. 2011, pp.263-267.

[102] A. Kousaridas, N. Alonistioti, “On a Synergetic Architecture for Cognitive Adaptive Behavior of Future Communication Systems”, In Proc IEEE WoWMoM, 2008, pp. 1-7.

[103] A. Mihailovic, I.P. Chochliouros, A. Kousaridas, G. Nguengang, et al. “Architectural Principles for Synergy of Self-management and Future Internet Evolution”, Proc. ICT Mobile and Wireless Communication Summit, Santander, Spain, 2009.

[104] A. Kousaridas, C. Polychronopoulos, N. Alonistioti, A. Marikar, J. Mödeker, A. Mihailovic, G. Agapiou, I. Chochliouros, G. Heliotis, “Future Internet elements: cognition and self-management design issues”, In Proc. ICST/ACM 2nd International Conference on Autonomic Computing and Communication Systems, 2008.

[105] T. Raptis, C. Polychronopoulos, A. Kousaridas, P. Spapis, V. Gazis, N. Alonistioti, I. Chochliouros, "Technological Enablers for Self-Manageable Future Internet Elements", In Proc. COMPUTATIONWORLD, 2009, pp.499-504.

[106] I.P. Chochliouros, E. Sfakianakis, A. Kousaridas, J. Modeker, D. Wagner, A.S. Spiliopoulou, G. Agapiou, et al. “Design and Development of Essential Use-Cases for Self-Management in Future Internet Wireless Networks”, In Proc. MOBILIGHT, 2010.

[107] A. Mihailovic, G. Nguengang, A. Kousaridas, M. Israel, V. Conan, et al., "An approach for designing cognitive self-managed Future Internet," In Proc. Future Network and Mobile Summit, Jun 2010, pp.1-9.

[108] M. Bouet, G. Nguengang, V. Conan, A. Kousaridas, P. Spapis, N. Alonistioti, “Embedding Cognition in the Wireless Network Management: An experimental perspective” IEEE Communications Magazine, vol. 50, no. 12, pp. 150-160, 2012.

[109] F. Cheong and R. Lai, “Designing a hierarchical fuzzy logic controller using the differential evolution approach”, Elsevier J. Applied Soft Computing, vol. 7, no. 2, March 2007, pp. 481-491.

[110] O. Sallent, J. Perez-Romero, R. Ljung, et al., “Operator's RAT Selection Policies Based on the Fittingness Factor Concept”, Proc. 16th IST Mobile and Wireless Communications Summit, Jul. 2007, pp. 1-5.

[111] A. Kousaridas, N. Alonistioti, A. Mihailovic, “Dynamic compartment formation for coverage optimization of cognitive wireless networks”, in Proc. IEEE PIMRC, Istanbul, 2010, pp. 2255-2260.

[112] A. Mihailovic, A. Kousaridas, A. Jaron, P. Pangalos, N. Alonistioti, H. A. Aghvami, “Self-Management of Future Access Networks for dynamic Configuration and Optimisation” Springer Wireless Personal Communications, 2013

[113] E. Schaeffer, “Graph clustering”, J. Computer Science Review, vol. 1, no. 1, pp. 27-64, 2007.

[114] J.Y. Yu, P.H.J Chong, "A survey of clustering schemes for mobile ad hoc networks," IEEE Trans. Commun. Surveys & Tutorials, vol. 7, no. 1, 2005, pp. 32-48.

[115] M.E.J. Newman, M. Girvan, “Finding and evaluating community structure in networks”, J. Phys. Rev. E 69, 026113, 2004.

[116] D.J., Baker, A. Ephremides, “The Architectural Organisation of a Mobile Radio Network Via a Distributed Algorithm”, IEEE Trans. Commun.,vol. 29, no. 11, pp. 1694-1701, 1981.

[117] A. Ephremides, J.E. Wieselthier, D.J. Baker, ”A design concept for reliable mobile radio networks with frequency hoping signaling”, in Proc. IEEE, vol. 75, pp. 56-73, 1987.

[118] P. Basu, N. Khan, T.D.C. Little, “A mobility based metric for clustering in mobile ad hoc networks”, in Proc. IEEE ICDCS, 2001, pp.413-418.

[119] A.K. Parekh, “Selecting routers in ad hoc wireless network”, Proc. SBT/IEEE ITS, 1994.

[120] M. Chatterjee, S. K. Das, and D. Turgut, “An On-Demand Weighted Clustering Algorithm (WCA) for Ad hoc Networks,” in Proc. IEEE Globecom, 2000, pp. 1697-701.

[121] S. Basagni, ‘Distributed clustering for ad hoc networks” in Proc. ISPAN, 1999, pp. 310–315.

[122] S. Basagni, “Distributed and mobility-adaptive clustering for multimedia support in multi-hop wireless networks”, in Proc. VTC, 1999, pp. 889-893.

[123] C.-C. Chiang, H.-K. Wu , W. Liu, et al., “Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel”, in Proc. IEEE SICON, 1997.

[124] S. Banerjee, S. Khuller, “A clustering scheme for hierarchical control in multi-hop wireless networks”, in Proc, IEEE INFOCOM 2001, Anchorage ,Alaska, April 2001, pp.1028-1037.

[125] H. Chan and A. Perrig, “ACE: an emergent algorithm for highly uniformcuster formation”, Lecture Notes in Computer Science, vol. 2920, pp. 154-171, 2004

[126] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks”, in Proc. Annual Hawaii International Conference, 2000.

[127] O. Younis, S. Fahmy, “HEED: A hybrid, energy-efficient, distributed clustering approach for Ad Hoc sensor networks”, IEEE Trans. Mobile Computing, vol. 3, no. 4, pp. 366-379, 2004.

[128] S. Yi, J. Heo, Y. Cho, and J. Hong, “PEACH: power-efficient and adaptive clustering hierarchy protocol for wireless sensor networks”, J. Computer Communications, vol. 30, no. 14-15, pp. 2842-2852, 2007.

[129] J. Congfeng, Y. Daomin, Z. Yinghui, "Towards Clustering Algorithms in Wireless Sensor Networks-A Survey," in Proc. IEEE WCNC, 2009, pp.1-6.

[130] S. Chinara, and S, K, Rath, “A Survey on One-Hop Clustering Algorithms in Mobile Ad Hoc Networks” J. Netw. Syst. Management, vol. 17, no. 1-2, pp. 183-207, 2009.

[131] M. E. J. Newman, “Modularity and community structure in networks”, in Proc. Natl. Acad. Sci. USA 103, pp. 8577–8582, 2006.

[132] A.-L. Barabási, Z. N. Oltvai “Network biology: understanding the cell’s functional organization”, Nature Reviews Genetics 5, pp. 101-113, 2004.

[133] M. E. J. Newman, and M. Girvan, “Finding and evaluating community structure in networks”, Phys. Rev. E 69(026113), pp. 56–68, 2004.

[134] T. Schank and D. Wagner, “Approximating clustering coefficient and transitivity”, J. Graph Algorithms and Appl., vol 9, no 2, pp. 265-275, Aug. 2005.

[135] NodeXL, http://nodexl.codeplex.com

[136] A. Kousaridas and N. Alonistioti, "Self-Organizing Cognitive Network Elements for Next Generation Communication System”, In Proc. NAEC, 2009.

[137] V. Gazis, A. Kousaridas, C. Polychronopoulos, T. Raptis, N. Alonistioti, "Self-Management Capacities in Future Internet Wireless Systems", In Proc. ADAPTIVE, 2009, pp. 9-14.

[138] A. Kousaridas and N. Alonistioti, "Topology Control in self-Managed Wireless Networks”, In Proc. MOBILIGHT, 2010.

[139] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “An Application-Specific Protocol Architecture for Wireless Microsensor Networks”, IEEE Trans. Wireless Commun., vol. 1, no. 4, pp. 660-670, Oct. 2002.

[140] A. Ephremides, “Energy concerns in wireless networks”, IEEE Commun. Mag, vol. 9, no 4, pp. 48-59, 2002.

[141] J. C. Park and S. K. Kasera, “Expected data rate: an accurate high-throughput path metric for multi-hop wireless routing,” in Proc. IEEE SECON, California, USA, Sept. 2005, pp. 218–228.

[142] A.Mishra, S. Banerjee, W. A. Arbaugh, “Weighted coloring based channel assignment for WLANs”. ACM SIGMOBILE J. Mobile Comput. and Commun. Review, vol. 9, no 3, pp. 19-31, 2005.

[143] OPNET Modeler, http://www.opnet.com

[144] K. Kalliojarvi, J. Pihlaja, A, Richter, et. al, “Cognitive Control Radio (CCR) - Enabling coexistence in Heterogeneous Wireless Radio Network”, in Proc. ICT Mobile Summit, Santander, Spain, Jun. 2009.

[145] D. Heckerman, “The Certainty-Factor Model”, http://research.microsoft.com/en-us/um/people/heckerman/h92encyclopedia.pdf

[146] A. Kousaridas, G. Katsikas, N. Alonistioti, E. Piri, M. Palola and J. Makinen (2011), "Testing End-to-End Self-Management in a Wireless Future Internet Environment", In. John Domingue, Alex Galis, Anastasius Gavras, Theodore Zahariadis, Dave Lambert. The Future Internet (pp. 259-270), Berlin, Heidelberg: Springer-Verlag.

[147] P. Magdalinos, A. Kousaridas, P. Spapis, G. Katsikas, N. Alonistioti, “Enhancing a Fuzzy Logic Inference Engine through Machine Learning for a Self-Managed Network”, Springer MONET 16(4), pp. 475-489, 2011.

[148] P. Magdalinos, A. Kousaridas, P. Spapis, G.Katsikas, N. Alonistioti, “Feedback-based Learning for Self-Managed Network Elements”, IEEE/IFIP International Symposium on Integrated Networks Management, Integrated Network Management, pp. 666-669, 2011.

[149] Panlab/PII project, http://www.panlab.net

[150] ITU-T Recommendation Y.2011, “General principles and general reference model for Next Generation Networks”, 2004.

[151] S. Sengupta, M. Chatterjee, S. Ganguly, “Improving Quality of VoIP Streams over WiMax”, IEEE Trans. on Computers, vol. 57, no. 2, Feb. 2008, pp. 145-156.

[152] A. Haseeb, V. Tralli, “QoS Performance Analysis for VoIP Traffic in Heterogeneous Networks with WiMAX Access”, in Proc. Intl. Conf. on Advanced Communication Technology, Seoul, Feb. 2011, pp. 960–965.

[153] J. Pan, T. Lee, W. Lee and C. Tsao “Improving the R-score of an adaptive VoIP codec in IEEE 802.16 networks” Springer J. Telecommunication Systems, 2011, pp. 1–13.

[154] K. Pentikousis, E. Piri, J. Pinola, et al. "Empirical evaluation of VoIP aggregation over a fixed WiMAX testbed", in Proc. ICST TridentCom, Austria, 2008.

[155] A. Kousaridas, P. Magdalinos, N. Alonistioti (2011), “Service Provision Evolution in Self-Managed Future Internet Environments”, In. Le Al-Sakib Khan Pathan, Mukaddim Pathan, and Hae Young, Advancements in Distributed Computing and Internet Technologies: Trends and Issues (pp. 112-140), IGI Global.

[156] ITU-T, Y.2012, “Functional requirements and architecture of next generation networks”, 2010.

[157] K. Rosen, Discrete Mathematics and Its Applications. McGraw-Hill, 2007.

[158] Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques. MA, USA:Morgan Kaufmann Publishers, 2012.

[159] Airspan homepage: http://www.airspan.com

[160] VTT Converging Networks Laboratory, Oulu, Finland, http://www.cnl.fi

[161] Distributed Internet Traffic Generator, http://www.grid.unina.it/software/ITG/index.php

[162] IEEE Std. 802.16-2004, “IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems” Oct. 2004.

[163] WiMAX forum, http://www.wimaxforum.org

[164] ITU-T Recommendation G.107, “The E-Model, a computational model for use in transmission planning”, Aug. 2009.

[165] S. Sengupta, M. Chatterjee, S. Ganguly, “Improving Quality of VoIP Streams over WiMax”, IEEE Trans. on Computers, vol. 57, no. 2, Feb. 2008, pp. 145-156.

[166] TM Forum GB934 “Best practice: Voice over IP SLA Management, Application Note to the SLA Management Handbook”, 2009.