NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

PhD THESIS

Context-based Resource Management and Slicing for SDN-enabled 5G Smart, Connected Environments

Sokratis N. Barmpounakis

ATHENS
MAY 2018

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Διαχείριση με επίγνωση πλαισίου και τεμαχισμός δικτυακών πόρων, για έξυπνα, συνδεδεμένα 5G περιβάλλοντα βασισμένα σε SDN

Σωκράτης Ν. Μπαρμπουνάκης

ΑΘΗΝΑ
ΜΑΪΟΣ 2018

PhD THESIS

Context-based Resource Management and Slicing
for SDN-enabled 5G Smart, Connected Environments

Sokratis N. Barmpounakis

SUPERVISOR: Athanasia (Nancy) Alonistioti, Assistant Professor UoA

THREE-MEMBER ADVISORY COMMITTEE:

Athanasia (Nancy) Alonistioti, Assistant Professor UoA
Lazaros Merakos, Professor UoA
Ioannis Stavrakakis, Professor UoA

SEVEN-MEMBER EXAMINATION COMMITTEE

(Signature)

Athanasia (Nancy) Alonistioti,
Assistant Professor UoA

(Signature)

Lazaros Merakos,
Professor UoA

(Signature)

Ioannis Stavrakakis,
Professor UoA

(Υπογραφή)

Dimitrios Vergados,
Associate Professor UniPi

(Υπογραφή)

Alexandros Kaloxylos,
Assistant Professor UoP

(Signature)

Dimitrios Syvridis
Professor UoA

Stathes Hadjiefthymiades,
Associate Professor UoA

Examination Date 25/05/2018

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Διαχείριση με επίγνωση πλαισίου και τεμαχισμός δικτυακών πόρων, για έξυπνα, συνδεδεμένα 5G περιβάλλοντα βασισμένα σε SDN

Σωκράτης Ν. Μπαρμπουνάκης

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Αθανασία (Νάνσυ) Αλωνιστιώτη, Επίκουρη Καθηγήτρια ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:
Αθανασία (Νάνσυ) Αλωνιστιώτη, Επίκουρη Καθηγήτρια ΕΚΠΑ
Λάζαρος Μεράκος, Καθηγητής ΕΚΠΑ
Ιωάννης Σταυρακάκης, Καθηγητής ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

(Υπογραφή)

Αθανασία (Νάνσυ) Αλωνιστιώτη,
Επίκουρη Καθηγήτρια ΕΚΠΑ

(Υπογραφή)

Λάζαρος Μεράκος,
Καθηγητής ΕΚΠΑ

(Υπογραφή)

Ιωάννης Σταυρακάκης,
Καθηγητής ΕΚΠΑ

(Υπογραφή)

Δημήτριος Βέργαδος
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Πειραιά

(Υπογραφή)

Αλέξανδρος Καλόξυλος,
Επίκουρος Καθηγητής
Πανεπιστήμιο Πελοποννήσου

(Υπογραφή)

Δημήτριος Συβρίδης,
Καθηγητής ΕΚΠΑ

Ευστάθιος Χατζηευθυμιάδης,
Αναπληρωτής Καθηγητής ΕΚΠΑ

Ημερομηνία εξέτασης 25/05/2018

ABSTRACT

The fifth-generation (5G) mobile communication systems, which are expected to emerge in the forthcoming years, will address unprecedented demands in terms of system capacity, service latency and number of connected devices. Future 5G network ecosystems will comprise a plethora of 3GPP and non-3GGP Radio Access Technologies (RATs), such as Wi-Fi, 3G, 4G or LTE, Bluetooth, etc. Deployment scenarios envision a multi-layer combination of macro, micro and femto cells where multi-mode end devices, supporting diverse applications, are served by different technologies. Limitations previously posed by legacy generation systems need to be eliminated, paving the way to a new wave of services and overall experience for the user. As a result, the management of radio resources via mapping the end devices to the most appropriate access network becomes of paramount importance; the primary Radio Resource Management (RRM) mechanisms, i.e. cell selection/reselection, handover and call admission control will be able to offer extremely high Quality of Service (QoS) and Experience (QoE) to the users, towards the very demanding 5G use case requirements; this will be realised via an optimal association between the diverse end devices and the coexisting available access networks. Besides the user's perspective, the Mobile Network Operators (MNOs) will be able to take advantage of the maximum efficiency and utilization over the –already scarce- wireless resources. Intelligent optimizations, as well as cost and energy efficient solutions need to be introduced in 5G networks in order to promote a consistent, user-centred and all-dimensional information ecosystem.

This thesis primarily focuses on the radio resource management (RRM) from the perspective of the primary RAT and cell layer selection processes (i.e., cell (re)selection, handover, admission control); afterwards, it goes one step beyond, in order to link the RRM with one of the latest RRM optimization approaches, i.e. the Network Slicing, as introduced in Software Defined Networking (SDN)-enabled environments, which creates smaller, virtual "portions" of the network, adapted and optimized for specific services/requirements. As a first step, a comprehensive analysis for the existing solutions -as these are specified in 3GPP standards, research papers, and patents has taken place. This thesis initially identifies the links between the research community efforts, the industry implementations, as well as the standardization efforts, in an attempt to highlight realistic solution implementations, identify the main goals, advantages and shortcomings of these efforts. As will be shown, existing solutions attempt to balance between implementation simplicity and solution optimality. Thus, solutions are either simple to implement but achieve sub-optimal solutions or provide significant improvements but their complexity and the burden placed on the network components –in terms of processing, as well as signaling resources- renders them unattractive for a real-life deployment.

Towards this end, this thesis introduces a context-based radio resource management (RRM) framework, comprised of three distinct mechanisms: Two out of the three mechanisms exploit contextual information with the aim of optimising the resource management and UE flows-RAT mapping, while the third mechanism acts with an augmenting role to the former two, by pre-processing the contextual information required by such, context-based mechanisms and –thus- by limiting the signaling cost required for communicating this contextual information among network entities. In addition to the three mechanisms, comprehensive analysis has taken place in relation to architectural aspects, in the context of the forthcoming 5G network architecture and by mapping them with the latest 5G network components –as these were introduced in the latest 3GPP work-.

The first major contribution of this thesis is COmpAsS, a context-aware, multi-criteria RAT selection mechanism, the main part of which operates on the User Equipment (UE) side, minimizing signaling overhead over the air interface and computation load on the base

stations. COmpAsS mechanism performs real-time monitoring and exploits the Fuzzy Logic (FL) approach as the core logic component, responsible for the perception of the network situation and -in combination with a set of pre-defined rules, calculates a list of the most suitable available access network options, for each one of the UE's active data flows/services. The merits of COmpAsS are showcased via an extensive series of simulation scenarios, as part of 5G ultra dense networks (UDN) use cases. The results prove how the proposed mechanism optimises Key Performance Indicators (KPIs), when juxtaposed to a well-established LTE handover algorithm.

The second major contribution of the current thesis the Context Extraction and Profiling Engine (CEPE), a resource management framework, which analyzes user behavioral patterns, extracts meaningful knowledge and performs user profiling in order to apply it for optimal resource planning, as well as prediction of resource requirements. CEPE collects information about users, services, terminals and network conditions and –based on offline processing– derives a knowledge model, which is subsequently used for the optimization of the primary RRM mechanism. Then, the extracted context information is translated into user profiles and is finally applied as input for enhanced cell (re)selection, handover or admission control. The viability and validity of CEPE is demonstrated via an extensive set of simulation scenarios.

The third major contribution is CIP, a Context Information Pre-processing scheme, aiming to identify and discard redundant or unnecessary data prior to network signaling and targeting to reduce the data used for knowledge extraction. CIP could be considered as an integral part of the afore described profiling schemes, i.e. COmpAsS and CEPE. The module comprises aggregating and compressing mobile network-related context information per unique identifier, such as the end device's International Mobile Subscriber Identity (IMSI), as well as techniques related to identifying and discarding user profile-redundant or unnecessary context data, before any transmission to CEPE. CIP gains are illustrated via a detailed analytical approach, guided by well-established 5G use case requirements.

As a final major contribution of this thesis, a comprehensive analysis takes place with regard to the CEPE-COmpAsS interworking, in the context of the forthcoming 5G network architecture and by mapping them with the latest 5G network components –as these were introduced in the latest 3GPP work-. The work in this section shows how the proposed framework can be instantiated as part of the 5G network components and functions introduced in SDN-enabled environments, such as the Network Slicing approach, the Network Data Analytics and the Network Slice Selection Functions, towards further optimising the distribution and management of the available infrastructure and network resources among the UEs, as well as the Access Traffic Steering, Switching and Splitting (ATSSS), responsible for managing the UE data flows and mapping each single UE flow with the optimal available access technology.

Two supplementary studies are finally included in this dissertation: a preliminary analysis on traffic engineering policies based on user profiling realised by CEPE, as well as a 5G use case related to the Internet of Things domain -and more specifically, Precision Farming-, aiming to highlight explicit requirements such as mission-critical machine type communication.

SUBJECT AREA: Communication Networks

KEYWORDS: 5G, Radio Resource Management, RAT selection, User Profiling, Handover, Context, SDN, Network Slicing

ΠΕΡΙΛΗΨΗ

Τα συστήματα κινητής επικοινωνίας πέμπτης γενιάς (5G) τα οποία αναμένονται τα αμέσως επόμενα χρόνια, θα αντιμετωπίσουν πρωτοφανείς απαιτήσεις όσον αφορά τον όγκο και το ρυθμό μεταδόσης δεδομένων, τις καθυστερήσεις του δικτύου, καθώς και τον αριθμό των συνδεδεμένων συσκευών. Τα μελλοντικά δικτυακά οικοσυστήματα θα περιλαμβάνουν μια πληθώρα τεχνολογιών ασύρματης επικοινωνίας (είτε τεχνολογιών 3GPP, είτε μη-3GPP) όπως το Wi-Fi, το 3G, το 4G ή LTE, το Bluetooth, κτλ. Τα σενάρια ανάπτυξης του 5G προβλέπουν έναν πολυεπίπεδο συνδυασμό μακρο- και μικρο-κυψελών, όπου πολυλειτουργικές συσκευές –οι οποίες μπορούν να υποστηρίξουν ποικιλία διαφορετικών εφαρμογών και υπηρεσιών- εξυπηρετούνται από διαφορετικές τεχνολογίες. Οι περιορισμοί που υπήρξαν στα παλιότερα συστήματα κινητών επικοινωνιών πρέπει να εξαλειφθούν, ανοίγοντας το δρόμο για ένα νέο κύμα υπηρεσιών και συνολική εμπειρία χρήστη. Ως εκ τούτου, η διαχείριση των ασύρματων πόρων μέσω της χαρτογράφησης και διανομής τους στις κινητές συσκευές, μέσω της πλέον κατάλληλης τεχνολογίας πρόσβασης, η οποία εξυπηρετεί τις ανάγκες των συγκεκριμένων υπηρεσιών/εφαρμογών αποκτά πρωταρχική σημασία. Οι κύριοι μηχανισμοί διαχείρισης πόρων δικτύου πρόσβασης δηλαδή η επιλογή κυψέλης (cell selection/reselection), η παράδοση υπηρεσίας από τη μία κυψέλη στην άλλη (handover), καθώς και ο έλεγχος εισαγωγής κλήσεων/υπηρεσιών (call/service admission control), είναι αυτοί που τελικώς θα μπορέσουν να προσφέρουν στους χρήστες εξαιρετικά υψηλή ποιότητα υπηρεσιών (Quality of Service - QoS) και εμπειρίας (Quality of Experience - QoE) προς τις πολύ απαιτητικές περιπτώσεις χρήσης του 5G. Αυτό θα γίνει εφικτό μέσω της βελτιστοποίησης του συσχετισμού-χαρτογράφησης μεταξύ των διαφορετικών (τελικών) κινητών συσκευών και των συνυπαρχόντων ασύρματων δικτύων πρόσβασης. Επιπλέον της οπτικής του χρήστη, οι Πάροχοι Δικτύων Κινητής θα είναι σε θέση να εκμεταλλευτούν τη μέγιστη αποδοτικότητα και χρήση των –ήδη δυσεύρετων- ασύρματων πόρων. Ευφυείς βελτιστοποιήσεις και αποδοτικές λύσεις όσον αφορά το κόστος και την κατανάλωση ενέργειας πρέπει επίσης να εισαχθούν στα δίκτυα 5ης γενιάς με σκοπό να προάγουν ένα συνεκτικό, στοχευμένο στο χρήστη και πολυδιάστατο οικοσύστημα πληροφοριών.

Η παρούσα διατριβή αυτή εστιάζει στη Διαχείριση Ασύρματων Δικτυακών Πόρων (ΔΑΔΠ - RRM) από την οπτική των κύριων διαδικασιών που σχετίζονται με την επιλογή ασύρματης τεχνολογίας πρόσβασης και στρώματος κυψέλης (μικρο-, μάκρο κυψέλη, κτλ.), δηλαδή η επιλογή κυψέλης, η παράδοση υπηρεσίας και ο έλεγχος εισαγωγής κλήσεων/υπηρεσιών. Έπειτα, η διατριβή προχωρά ένα βήμα παραπέρα, με σκοπό να συνδέσει τη ΔΑΔΠ με μία από τις πιο πρόσφατες προσεγγίσεις διαχείρισης δικτυακών πόρων, δηλαδή τον «τεμαχισμό δικτύου» (network slicing), όπως αυτή εισάγεται σε περιβάλλοντα που χρησιμοποιούν τη μέθοδος της Δικτύωσης Βασισμένης στο Λογισμικό (Software Defined Networking), η οποία δημιουργεί μικρότερα, εικονικά τμήματα του δικτύου, προσαρμοσμένα και βελτιστοποιημένα για συκεκριμένες υπηρεσίες και αντίστοιχες απαιτήσεις. Σαν πρώτο βήμα, πραγματοποιήθηκε μια ολοκληρωμένη ανάλυση για τις υπάρχουσες λύσεις – όπως αυτές προδιαγράφονται στα πρότυπα της 3GPP, στη βιβλιογραφία, καθώς και τις σχετικές πατέντες-. Η διατριβή αυτή αρχικά εντοπίζει τους δεσμούς μεταξύ των προσπαθειών της ερευνητικής κοινότητας, των υλοποιήσεων της βιομηχανίας, καθώς και των δράσεων προτυποποίησης, σε μια προσπάθεια να επισημανθούν ρεαλιστικές λύσεις εφαρμογής, να προσδιοριστούν οι κύριοι στόχοι, τα πλεονεκτήματα, αλλά και οι ελλείψεις αυτών των προσπαθειών. Όπως θα δειχθεί, οι υπάρχουσες λύσεις προσπαθούν να εξισορροπήσουν σε ένα σημείο μεταξύ της βέλτιστης λύσης και μιας απλής υλοποίησης. Έτσι, οι λύσεις που έχουν προταθεί είτε είναι απλοποιημένες σε τέτοιο βαθμό που απομακρύνονται από μια ρεαλιστική πρόταση, και επιτυγχάνουν υπο-βέλτιστες λύσεις ή από την άλλη παρέχουν πολύ σημαντικές βελτιώσεις, αλλά η πολυπλοκότητά τους και η επιβάρυνση που επιβάλλουν στο δίκτυο

(όσον αφορά για παράδειγμα κόστος σηματοδοσίας, ή επεξεργαστικής ισχύος) τις καθιστούν ελκυστικές για μια πραγματική ανάπτυξη.

Προς αυτή την κατεύθυνση, η παρούσα διατριβή εισαγωγή ένα σύνολο μηχανισμών επίγνωσης πλαισίου για τη διαχείριση δικτυακών πόρων, που αποτελείται από τρεις επιμέρους μηχανισμούς με διακριτό ρόλο: Δύο από τους μηχανισμούς χρησιμοποιούν πληροφορία πλαισίου με σκοπό τη βελτίωση τη διαχείριση πόρων και και τη χαρτογράφηση μεταξύ ροών δεδομένων κινητών συσκευών και κυψέλης/τεχνολογίας δικτύου. Ο τρίτος μηχανισμός δρα με έναν ενισχυτικό ρόλο στους δύο προηγούμενους, μέσω μιας προ-επεξεργασίας που πραγματοποιεί πάνω σε πληροφορία πλαισίου, με σκοπό τον περιορισμό του κόστους της επιπλέον σηματοδοσίας που απαιτείται για την μεταφορά της πληροφορίας πλαισίου μεταξύ των διαφόρων ενδιαφερόμενων δικτυακών οντοτήτων. Εκτός από τους τρεις μηχανισμούς αυτούς, πραγματοποιήθηκαν εκτενείς μελέτες σε σχέση με αρχιτεκτονικά ζητήματα και πτυχές, στο πλαίσιο της επικείμενης αρχιτεκτονικής δικτύου 5G και χαρτογράφηση των προτεινόμενων μηχανισμών στα συστατικά στοιχεία του δικτύου 5G -όπως αυτά εισήχθησαν στα τελευταίο κείμενα προτυποποίησης της 3GPP-.

Η πρώτη κύρια συμβολή της παρούσας διατριβής είναι το COmpAsS, ένας μηχανισμός επιλογής Τεχνολογίας Ασύρματης Πρόσβασης πολλαπλών κριτηρίων, με γνώμονα το περιβάλλον, το κύριο μέρος του οποίου λειτουργεί στην πλευρά του Εξοπλισμού Χρήστη (UE), ελαχιστοποιώντας με αυτό τον τρόπο τις επιβαρύνσεις σηματοδότησης στη διεπαφή αέρα και το φορτίο υπολογισμού στους σταθμούς βάσης. Ο μηχανισμός COmpAsS εκτελεί παρακολούθηση σε πραγματικό χρόνο, υιοθετώντας την Ασαφή Λογική (Fuzzy Logic -FL) ως μία από τις βασικές προσεγγίσεις αντίληψης και ανάλυσης της κατάστασης του δικτύου. Σε συνδυασμό με ένα σύνολο προκαθορισμένων κανόνων, υπολογίζει μια λίστα με τις καταλληλότερες διαθέσιμες επιλογές πρόσβασης δικτύου, για κάθε μία από τις ροές δεδομένων/υπηρεσίας που είναι ενεργές εκείνη τη στιγμή. Τα πλεονεκτήματα του COmpAsS παρουσιάζονται μέσω μιας εκτεταμένης σειράς σεναρίων προσομοίωσης, ως μέρος των περιπτώσεων χρήσης εξαιρετικά πυκνών δικτύων (UDN) 5G. Τα αποτελέσματα αποδεικνύουν τον τρόπο με τον οποίο ο προτεινόμενος μηχανισμός βελτιστοποιεί τους βασικούς δείκτες επιδόσεων (Key Performance Indicators - KPIs), όταν αντιπαρατίθεται σε έναν από τους καθιερωμένους LTE αλγορίθμους.

Η δεύτερη σημαντική συμβολή της παρούσας διατριβής είναι η Μηχανή Εξόρυξης Πλαισίου και Δημιουργίας Προφίλ (Context Extraction and Profiling Engine – CEPE), ένας μηχανισμός διαχείρισης πόρων, ο οποίος αναλύει συμπεριφορικά πρότυπα των χρηστών/κινητών συσκευών, εξάγει ουσιώδη γνώση και δημιουργεί αντίστοιχα προφίλ/πρότυπα συμπεριφοράς, με σκοπό να τα χρησιμοποιήσει για βέλτιστο προγραμματισμό πόρων, καθώς επίσης και για την μελλοντική πρόβλεψη απαιτήσεων πόρων. Το CEPE συλλέγει πληροφορίες σχετικά με τους χρήστες, τις υπηρεσίες, τις κινητές συσκευές, καθώς και τις συνθήκες δικτύου, και μέσω επεξεργασίας -χωρίς σύνδεση, ετεροχρονισμένα- αποκτά ένα μοντέλο γνώσης, το οποίο στη συνέχεια χρησιμοποιείται για τη βελτιστοποίηση των κύριων μηχανισμών ΔΑΔΠ (RRM). Το προαναφερθέν μοντέλο γνώσης μεταφράζεται έπειτα σε προφίλ χρηστών/κινητών συσκευών, τα οποία εφαρμόζονται ως είσοδος κατά τις διαδικασίες ΔΑΔΠ. Η βιωσιμότητα και η εγκυρότητα του CEPE επιδεικνύεται μέσω εκτεταμένων σεναρίων προσομοίωσης.

Η τρίτη σημαντική συμβολή είναι το CIP (Context Information Preprocessor), ένας μηχανισμός προεπεξεργασίας πληροφοριών πλαισίου, με στόχο τον εντοπισμό και την απόρριψη περιττών δεδομένων κατά τη σηματοδοσία πριν από την εξαγωγή της γνώσης. Το CIP θα μπορούσε να θεωρηθεί ως αναπόσπαστο μέρος των προαναφερθέντων σχημάτων σχεδίασης, δηλαδή των COmpAsS και CEPE. Ο προτεινόμενος μηχανισμός περιλαμβάνει τη συγκέντρωση και συμπίεση πληροφοριών πλαισίου σχετικά με το δίκτυο ανά μοναδικό αναγνωριστικό κινητής συσκευής/χρήστη, -όπως η διεθνής ταυτότητα

συνδρομητή κινητού (IMSI)-, καθώς και τεχνικές που σχετίζονται με την αναγνώριση και την απόρριψη δεδομένων πλαισίου που δε συμβάλλουν στην βελτίωση ή διόρθωση του πρόφιλ χρήστη, πριν από οποιαδήποτε μετάδοση προς το CEPE (ή άλλο μηχανισμό ΔΑΔΠ). Οι βελτιώσεις και τα κέρδη του CIP στη διαδικασία της σηματοδοσίας απεικονίζονται μέσω λεπτομερούς αναλυτικής προσέγγισης, η οποία καθορίζεται από τις καθιερωμένες απαιτήσεις περί χρήσης 5G.

Ως τελική σημαντική συμβολή αυτής της διατριβής, διεξάγεται μια εκτεταμένη ανάλυση όσον αφορά τη διασύνδεση των CEPE-COmpAsS, στο πλαίσιο της επικείμενης αρχιτεκτονικής δικτύου 5G και της χαρτογράφησης αυτών με τα τελευταία συστατικά στοιχεία του δικτύου 5G –όπως αυτά παρουσιάστηκαν στις τελευταίες δημοσιεύσεις προτυποποίησης της 3GPP -. Το έργο σε αυτή την ενότητα δείχνει πώς μπορεί να παρουσιαστεί το προτεινόμενο πλαίσιο ως μέρος των συνιστωσών του δικτύου 5G και των λειτουργιών που εισάγονται σε περιβάλλοντα με δυνατότητα SDN, όπως η προσέγγιση του «Τεμαχισμού Δικτύου», ο Μηχανισμός Ανάλυσης Δικτυακών Δεδομένων (Network Data Analytics Function – NWDAF), η λειτουργία επιλογής βέλτιστου τεμαχίου δικτύου (Network Slice Selection Function) - προς περαιτέρω βελτιστοποίηση της διανομής και της διαχείρισης των διαθέσιμων πόρων δικτύου μεταξύ των συσκευών-, καθώς και το ATSSS – Access Traffic Steering, Switching and Splitting, μια οντότητα υπεύθυνη για τη διαχείριση των ροών δεδομένων των UE –με δυνατότητες επαναδρομολόγησης, διαχωρισμού και σύνδεσης της κάθε ροής με την αντίστοιχη βέλτιστη, διαθέσιμη τεχνολογία πρόσβασης.

Δύο συμπληρωματικές μελέτες περιλαμβάνονται –τέλος- σε αυτή τη διατριβή: μια αρχική ανάλυση των πολιτικών μηχανικής κυκλοφορίας (Traffic Engineering) που βασίζονται σε προφίλ χρηστών που προκύπτουν από το CEPE, καθώς και μία περίπτωση χρήσης 5G που σχετίζεται με τον τομέα του Διαδικτύου των Πραγμάτων - και πιο συγκεκριμένα την «Καλλιέργεια Ακριβείας» (Precision Farming), με σκοπό να δοθεί έμφαση σε ρητές απαιτήσεις των περιπτώσεων χρήσης 5G, όπως η επικοινωνία τύπου μηχανής κρίσιμης σημασίας (Mission-Critical Machine Type Communication).

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα Επικοινωνιών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Δίκτυα 5ης Γενιάς, Διαχείριση Πόρων Ασύρματου Δικτύου, Επιλογή Ασύρματης Τεχνολογίας Πρόσβασης, Προφίλ Χρηστών, Παράδοση Υπηρεσίας, Πλαίσιο, Δικτύωση Βασισμένη στο Λογισμικό, Τεμαχισμός Δικτύου

ACKNOWLEDGEMENTS

At the outset, I would like to express my sincere gratitude to my thesis advisory committee: Assistant Professor Nancy Alonistioti, Professor Lazaros Merakos and Professor Ioannis Stavrakakis.

I would like to thank and express my appreciation to my advisor, Assistant Professor Nancy Alonistioti for supporting me and guiding me throughout all these years. Our academic interactions have always been beneficial and thought provoking. I would also like to thank you for all the professional experiences and opportunities that I acquired because of you, as well as the excellent collaboration that we had. Your support, ideas, and feedback have been absolutely invaluable for this thesis and my research activities.

I want to express my gratitude to Assistant Professor Alexandros Kaloxylos, for his invaluable support and guidance since the very beginning of this thesis. Alex's vision and inputs have been some of the cornerstones in this work and without his help, perhaps this thesis would not exist. You are a very inspiring and mind-triggering guy, -not only from the professional point of view, but as a person, as a whole-, and our discussions have influenced several aspects of the way I see things in life.

I would like to sincerely thank my friends and ex colleagues Dr. Panagis Magdalinos and Dr. Panagiotis Spapis for their invaluable help and contribution, during our collaboration, in parts of this thesis. Your expertise and deep knowledge helped me in numerous parts of this work. I would like to thank my friends and (ex or present) colleagues Dr. Konstantinos Chatzikokolakis, Dimitris Soukaras, Stamatis Kolovos, Makis Stamatelatos, Dr. Apostolos Kousaridas, Aggelos Groumas, Lampros Katsikas, Konstantina Dimtsa, George Kyprianidis, Roi Arapoglou, Stefanos Falangitis, George Beinas, Panagiotis Panagiotopoulos, Vassilis Sarris, Michail Papadakis and Stella Kazilieri. You have all directly or indirectly contributed in the shaping of this thesis and my personal research profile.

I want to thank my father, Nikos, for being the root and inspiration of my scientific journey since I was a kid. He is the one who made me see the beauty in exploring engineering and science and taught me to always target more in life. His way of thinking made me a better human being, and I thank him for that. I want to thank my brother, (ex-)roommate and friend, Dr. Manos Barmpounakis for his love, support and endless, fruitful, scientific and philosophical discussions we had. Also, his devotion and hard work during his own PhD journey was a great inspiration for me. I want to thank my mother, Mary for inspiring me with her strength, her persistence and positive way of facing challenges in life. I want to thank my younger sister, Irini for her love and support. I want to express my love and appreciation to Anastasia, my life partner, for supporting me, inspiring me with her beautiful worldview and providing me with strength and courage during all these years.

Last but not least, I would like to thank my friends, and fellow musicians in our band, "astrogono" («αστρογόνο»), Orestis Tsichlakis, Vassilis Kourtis, Giannis Mavritsakis, Stefanos Mouroutsos, Thanos Athanasopoulos and Thomas Meleteas for being there, exploring together our music journey and creating together some of the most meaningful moments of our lives. Music, -which is no hobby for me, but one of the deepest and most meaningful elements of my life-, fills the other half of myself, next to science and engineering, and without it, nothing would be the same.

Στο Δημήτρη

PUBLICATIONS LIST

Refereed Journal Publications

1. **S. Barmpounakis**, A. Kaloxylos, P. Spapis, N. Alonistioti, «Context-aware, user-driven, network-controlled RAT selection for 5G networks», The International Journal of Computer and Telecommunications Networking (COMNET Journal), Elsevier, Vol. 113, pp. 124-147, February 2017.
2. D. Calabuig, **S. Barmpounakis**, et al., «Resource and Mobility Management in the Network Layer of 5G Cellular Ultra-Dense Networks», IEEE Communications Magazine, Vol. 55, Issue: 6, pp. 162-169, 2017.
3. P. Magdalinos, **S. Barmpounakis**, P. Spapis, A. Kaloxylos, et al., «A Context Extraction and Profiling Engine for 5G Network Resource Mapping», The International Journal for the Computer and Telecommunications Industry (COMCOM Journal), Elsevier Computer Communications, Volume 109, pp. 184-201, September 2017
4. **S. Barmpounakis**, P. Spapis, A. Kaloxylos, P. Magdalinos, N. Alonistioti, C. Zhou, "RAT and cell layer selection in 5G networks: Mechanisms and Trends - User Profiling and Data Analytics towards Radio Resource Management functions optimization", IEEE Communications Surveys & Tutorials *(submitted, undergoing revisions)*
5. **S. Barmpounakis**, P. Spapis, A. Kaloxylos, P. Magdalinos, M. Stamatelatos, N. Alonistioti, C. Zhou, "A Resource Management Framework Based on User Profiling and Signaling Minimization for 5G Networks" *(submitted, undergoing revisions)*
6. **S. Barmpounakis**, A. Kaloxylos, A. Groumas et al., «Management & Control applications in Agriculture domain via a Future Internet Business-to-Business Platform», Information Processing in Agriculture Journal (IPA), Vol. 2, Issue 1, pp. 51-63, May 2015

Patents

1. **WO2017137089A1**, World Intellectual Property Organization, EP 2016/052963, European Patent Office, "User Equipment Profiling for Network Administration", A. Kaloxylos, P.Spapis, C. Zhou, **S. Barmpounakis**, N. Alonisitioti
2. **WO2017211415A1**, World Intellectual Property Organization, EP 2016/063081, European Patent Office, "Context Information Processor, Profile Distribution Unit and Method for a Communication Network", A. Kaloxylos, P.Spapis, C. Zhou, **S. Barmpounakis**, N. Alonisitioti

Referred Conference Publications

1. **S. Barmpounakis**, A. Kaloxylos, P. Spapis, N. Alonistioti, "COmpAsS: A Context-Aware, User-Oriented RAT Selection Mechanism in Heterogeneous Wireless Networks", Mobility 2014, Fourth International Confrence on Mobile Services, Resources, and Users, July 20-24 – 2014, Paris, France
2. A. Kaloxylos, **S. Barmpounakis**, P. Spapis, N. Alonistioti, "An efficient RAT selection mechanism for 5G cellular networks", International Wireless

Communications and Mobile Computing Conference, 4-8 August 2014, Nicosia, Cyprus

Whitepapers

1. 5G-PPP, “5G Automotive Vision”, October 2015, link [July 2017]: *https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf*

ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Στα πλαίσια της παρούσας διατριβής σχεδιάζεται ένας συνολικός Μηχανισμός διαχείρισης πόρων ασύρματων δικτύων σε περιβάλλοντα που αποτελούνται από ετερογενείς τεχνολογίες ασύρματης πρόσβασης. Τέτοια περιβάλλοντα είναι τα δικτυακά περιβάλλοντα όπως περιγράφονται σε σενάρια για τα επερχόμενα συστήματα κινητής επικοινωνίας πέμπτης γενιάς (5G). Ο εν λόγω Μηχανισμός αποτελείται από επιμέρους μηχανισμούς που έχουν την δυνατότητα να λειτουργήσουν είτε σαν αυτοτελείς οντότητες και να βελτιώσουν συγκεκριμένα στοιχεία της λειτουργίας του δικτύου, είτε συμπληρωματικά για μια ολιστική αντιμετώπιση του προβλήματος της διαχείρισης των ασύρματων πόρων σε ετερογενή δικτυακά περιβάλλοντα 5ης γενιάς. Ο Μηχανισμός αποτελεί ένα σχήμα με το οποίο μπορεί ένα ή περισσότερα δικτυακά στοιχεία να μοντελοποιούν το περιβάλλον τους και να παίρνουν αποφάσεις για την κατάσταση στην οποία βρίσκονται, καθώς και τη βελτιστοποίηση της διανομής των ασύρματων πόρων στις κινητές συσκευές που τους καταναλώνουν, ακόμα και όταν αυτές έχουν διαφορετικές τελείως απαιτήσεις μεταξύ τους. Επιπλέον, χρησιμοποιούνται μηχανισμοί εκμάθησης ώστε σε ενδεχόμενη αλλαγή των συνθηκών του περιβάλλοντος, ο Μηχανισμός να έχει τη δυνατότητα να επαναξιολογήσει το περιβάλλον, να εξελίσσει το σχήμα αντίληψης τρέχουσας κατάστασης βασισμένο στην εξέλιξη των περιβάλλουσων συνθηκών, και με αυτόν τον τρόπο να απαλλαγεί το σύστημα από την ανάγκη της χειροκίνητης παραμετροποίησης, η οποία είναι χρονοβόρα και επιρρεπής σε σφάλματα.

Ένα σύστημα διαχείρισης δικτυακών πόρων ασύρματης πρόσβασης με επίγνωση πλαισίου ενσωματώνει όλες εκείνες τις λειτουργίες που του επιτρέπουν να παρατηρεί το περιβάλλον του, να αντιλαμβάνεται την κατάσταση στην οποία βρίσκεται, και στη συνέχεια να λαμβάνει τις βέλτιστες αποφάσεις για τη διαχείριση και διανομή των πόρων αυτών στις διαθέσιμες κινητές συσκευές.

Κύριο κομμάτι της λειτουργίας τέτοιων συστημάτων, είναι η δυνατότητά τους (και των αντίστοιχων δικτυακών στοιχείων) να παρατηρούν το περιβάλλον και να προχωρούν σε ανάλυση της κατάστασης, για να αναγνωρίσουν προβληματικές καταστάσεις ή ενδεχόμενες ευκαιρίες βελτιστοποίησης της λειτουργίας του συστήματος. Η παραπάνω διαδικασία αναφέρεται ως λειτουργικότητα επίγνωσης κατάστασης ή πλαισίου. Η συγκεκριμένη λειτουργία έχει προταθεί από διάφορους ερευνητές στη βιβλιογραφία. Στα πλαίσια της τρέχουσας διατριβής, αναλύθηκαν όλοι οι υπάρχοντες διαφορετικοί τρόποι αντιμετώπισης, καθώς και όλες τις αντιπροσωπευτικές προτάσεις της βιβλιογραφίας. Η λύση που εντάσσεται σε αυτή τη διατριβή παρουσιάζει σημαντικά πλεονεκτήματα –όπως αποδεικνύεται αναλυτικά στη συνέχεια- για τη συγκεκριμένη φύση των συστημάτων κινητής επικοινωνίας ετερογενών δικτύων, σε περιβάλλοντα 5^{ης} γενιάς.

Τα συστήματα κινητών επικοινωνιών πέμπτης γενιάς (5G), μέρος των οποίων έχουν ήδη ξεκινήσει να δοκιμάζονται λίγο πριν λειτουργήσουν για το κοινό, θα αντιμετωπίσουν πρωτοφανείς απαιτήσεις όσον αφορά τη χωρητικότητα του συστήματος, τους ρυθμούς μετάδοσης, τους χρόνους καθυστέρησης, καθώς και τον αριθμό των συνδεδεμένων συσκευών. Τα μελλοντικά συστήματα δικτύου 5G θα περιλαμβάνουν μια πληθώρα ετερογενών τεχνολογιών ραδιοεπικοινωνίας όπως το 3G, το 4G ή LTE, το 5G New Radio, το WiFi, το Bluetooth, κτλ. Τα σενάρια ανάπτυξης για τα συστήματα 5^{ης} γενιάς προβλέπουν συνδυασμό πολυστρωματικών μακροκυψελών και μικροκυψελών, όπου πολυλειτουργικές συσκευές, υποστηρίζοντας διαφορετικές εφαρμογές, εξυπηρετούνται από διαφορετικές τεχνολογίες ασύρματης πρόσβασης. Οι περιορισμοί που δημιουργήθηκαν από τα παλαιότερα συστήματα πρέπει να εξαλειφθούν, ανοίγοντας το δρόμο για ένα νέο κύμα υπηρεσιών και μια νέα συνολική εμπειρία για τον χρήστη.

Ως εκ τούτου, η διαχείριση των ασύρματων πόρων μέσω της χαρτογράφησης και διανομής τους στις (τελικές) κινητές συσκευές, μέσω της πλέον κατάλληλης τεχνολογίας πρόσβασης, η οποία εξυπηρετεί τις ανάγκες των συγκεκριμένων υπηρεσιών/εφαρμογών αποκτά πρωταρχική σημασία. Οι κύριοι μηχανισμοί διαχείρισης ραδιοπόρων (Radio Resource Management), δηλαδή η επιλογή κυψέλης (cell selection/reselection), η παράδοση υπηρεσίας από τη μία κυψέλη στην άλλη (handover), καθώς και ο έλεγχος εισαγωγής κλήσεων/υπηρεσιών (call/service admission control), είναι αυτοί που τελικώς θα μπορέσουν να προσφέρουν στους χρήστες εξαιρετικά υψηλή ποιότητα υπηρεσιών (Quality of Service - QoS) και εμπειρίας (Quality of Experience - QoE) προς τις πολύ απαιτητικές περιπτώσεις χρήσης του 5G. Αυτό θα πραγματοποιηθεί μέσω της και χαρτογράφησης (mapping), καθώς και βέλτισης αντιστοίχισης μεταξύ των διάφορων τελικών συσκευών και των συνυπάρχοντων διαθέσιμων δικτύων πρόσβασης. Παράλληλα με το χρήστη, οι πάροχοι συστημάτων κινητών δικτύων (Mobile Network Operators - MNOs) θα επωφεληθούν επίσης από τη μέγιστη αποδοτικότητα και αξιοποίηση των –δαπανηρών- ασύρματων πόρων. Θα πρέπει -ως εκ τούτου- να εισαχθούν ευφυείς βελτιστοποιήσεις, καθώς και αποδοτικές λύσεις κόστους και ενέργειας στα δίκτυα 5^{ης} γενιάς, προκειμένου να προωθηθεί ένα συνεκτικό οικοσύστημα εστιασμένο στο χρήστη, καθώς και όλους τους τύπους δεδομένων που μεταδίδονται.

Όπως προαναφέρθηκε, το σύστημα διαχείρισης που αναπτύχθηκε στα πλαίσια της διδακτορικής διατριβής αυτής, αποτελείται από τρεις επιμέρους μηχανισμούς. Ο πρώτος μηχανισμός είναι το *COmpAsS*, ένας μηχανισμός επιλογής Τεχνολογίας Ασύρματης Πρόσβασης πολλαπλών κριτηρίων με επίγνωση κατάστασης περιβάλλοντας που κάνει χρήση Ασαφούς Λογικής (Fuzzy Logic). Το κύριο μέρος του *COmpAsS* λειτουργεί στην πλευρά του Εξοπλισμού Χρήστη (UE), ελαχιστοποιώντας τις επιβαρύνσεις σηματοδότησης πάνω από τη διεπαφή αέρα και το φορτίο υπολογισμού στους σταθμούς βάσης. Ο μηχανισμός COmpAsS εκτελεί παρακολούθηση και καταγραφή των συνθηκών του δικτύου σε πραγματικό χρόνο και, σε συνδυασμό με ένα σύνολο προκαθορισμένων κανόνων –τους οποίους ορίζει ο εκάστοτε διαχειριστής του δικτύου-, υπολογίζει μια λίστα προτεραιότητας που απαρτίζεται από τις καταλληλότερες/διαθέσιμες επιλογές/τεχνολογίες πρόσβασης δικτύου. Τα πλεονεκτήματα του COmpAsS παρουσιάζονται μέσω μιας εκτεταμένης σειράς σεναρίων προσομοίωσης, ως μέρος των περιπτώσεων χρήσης εξαιρετικά πυκνών δικτύων 5G (Ultra Dense Networks). Τα αποτελέσματα αποδεικνύουν τον τρόπο με τον οποίο ο προτεινόμενος μηχανισμός βελτιστοποιεί τους βασικούς δείκτες επιδόσεων (Key Performance Indicators - KPIs), όταν αντιπαρατίθενται σε παραδοσιακούς μηχανισμούς του LTE. Ο δεύτερος μηχανισμός που προτείνεται αποτελεί μια μηχανή ανάλυσης πληροφορίας πλαισίου και εξόρυξης προφίλ χρηστών και ονομάζεται CEPE (Context Extraction and Profiling Engine). Το CEPE βασίζεται σε μηχανισμούς μηχανικής μάθησης για να αναλύσει την πληροφορία πλαισίου και περιβάλλοντος, τα πρότυπα συμπεριφοράς των χρηστών, να εξάγει σημαντικές γνώσεις και έπειτα εφαρμόσει τις μεθόδους διαχείρισης των πόρων του δικτύου βάσει αυτών των προτύπων συμπεριφοράς/προφίλ για βέλτιστο προγραμματισμό. Η εγκυρότητα του CEPE και η βελτιστοποίηση που προσφέρει στη διαχείριση των πόρων του δικτύου, επίσης αποδεικνύεται μέσα από μια σειρά προσομοιώσεων σε ρεαλιστικά και σύνθετα δικτυακά περιβάλλοντα 5^{ης} γενιάς. Ο τρίτος μηχανισμός είναι το CIP (Context Information Pre-Processing), ένα πρόγραμμα προ-επεξεργασίας πληροφοριών περιβάλλοντος/πλαισίου, με στόχο τον εντοπισμό και την απόρριψη περιττών ή επαναλαμβανόμενων δεδομένων πριν από την εξαγωγή της γνώσης. Ο τελικός στόχος του συγκεκριμένου μηχανισμού είναι η ελαχιστοποίηση της πληροφορίας που εντάσσεται στην σηματοδοσία επιπλέον πληροφορίας πλαισίου. Το CIP θα μπορούσε να θεωρηθεί ως αναπόσπαστο μέρος των προαναφερθέντων σχημάτων, δηλαδή τα COmpAsS και CEPE. Ο μηχανισμός του CIP περιλαμβάνει την συμπίεση πληροφοριών πλαισίου ανά μοναδικό αναγνωριστικό, όπως η διεθνής

ταυτότητα συνδρομητή κινητού (IMSI) της συσκευής, καθώς και τεχνικές προεπεξεργασίας και φιλτραρίσματος που σχετίζονται με την αναγνώριση και την απόρριψη περιττών πληροφοριών προφίλ χρήστη πριν από οποιαδήποτε μετάδοση προς το CEPE. Τα κέρδη του CIP σε σχέση με τη μείωση σηματοδοσίας δικτύου αποδεικνύονται μέσω λεπτομερούς αναλυτικής προσέγγισης, η οποία πλαισιώνεται από τις καθιερωμένες απαιτήσεις χρήσης 5G. Οι τρεις αυτοί μηχανισμοί που αναφέρθηκαν παραπάνω περιληπτικά, περιγράφονται στο τελευταίο μέρος της διατριβής ως μέρος των βασικών δομικών δικτυακών στοιχείων του συστήματος κινητών επικοινωνιών 5ης γενιάς, παρουσιάζεται μια εκτεταμένη συζήτηση για το ρόλο που επιτελεί το καθένα, καθώς και τις αναγκαίες προσαρμογές που απαιτούνται έτσι ώστε η λειτουργία τους να είναι συμβατή με ένα 5G δίκτυο. Επιπλέον, επεκτείνεται η RRM λειτουργικότητά τους με την ενσωμάτωση των τεχνολογιών της Δικτύωσης Βασισμένης στο Λογισμικό (Software Defined Networking) καθώς και της τεχνολογίας «Τεμαχισμού» Δικτυακών Πόρων (Network Slicing), κατά τις οποίες, μέρος των υποδομών και των διαθέσιμων πόρων του συστήματος, διατίθενται αποκλειστικά και δυναμικά για συγκεκριμένα είδη υπηρεσιών με καθορισμένες απαιτήσεις σε εύρος φάσματος, λειτουργίες δικτύου (network functions), επιτρεπόμενες καθυστερήσεις βάσει τις κρισιμότητας της υπηρεσίας, κτλ. Το συνολικό δίκτυο «τεμαχίζεται» με αυτό τον τρόπο σε μικρότερα υποδίκτυα, με προσαρμοσμένα χαρακτηριστικά, τα οποία βελτιστοποιούν την απόδοση συγκεκριμένων «κάθετων» εφαρμογών και τομέων («5G vertical domains»).

Η συγκεκριμένη διατριβή περιλαμβάνει εννέα κεφάλαια. Το πρώτο κεφάλαιο αποτελεί την εισαγωγή στο αντικείμενο της διατριβής. Σε αυτό το κεφάλαιο αποτυπώνονται οι τεχνολογικές εξελίξεις και οι προκλήσεις των μελλοντικών δικτύων επικοινωνιών. Οι προκλήσεις αυτές αφορούν κυρίως στη δυνατότητα των δικτυακών στοιχείων να λειτουργήσουν σε πολύ πυκνά δικτυακά περιβάλλοντα (UDNs)· η πυκνότητα σχετίζεται με τον αριθμό των δικτυακών συσκευών και την ενδεχόμενη διαλειτουργικότητα και συνεργασία τους με σκοπό να εξυπηρετούνται όλοι οι τερματικές (κινητές) συσκευές με ικανοποιητικό βαθμό υπηρεσίας (QoS). Πέρα από τον αριθμό των συσκευών που θα αυξηθεί κατακόρυφα και θα οδηγήσει στα πυκνά δικτυακά περιβάλλοντα 5ης γενιάς (Ultra Dense Networks – UDNs), αυξάνονται συνεχώς και οι απαιτήσεις όπως και ο όγκος της πληροφορίας που πρέπει να μεταδίδεται ανά εφαρμογή, εξαιτίας της αύξησης των προσφερόμενων υπηρεσιών (υψηλότερη ανάλυση βίντεο, μικρότερες καθυστερήσεις απόκρισης για κρίσιμες υπηρεσίες όπως υπηρεσίες συνδεδεμένων «ημι-αυτόνομων» οχημάτων, κτλ.). Επιπλέον, τα πολύ πυκνά δίκτυα UDNs χαρακτηρίζονται από μεγάλη πυκνότητα εγκατεστημένων σταθμών βάσης, οι οποίοι μπορούν να ποικίλουν σε είδος, ανάλογα με το εύρος και την ισχύ εκπομπής τους: Πέρα από τους παραδοσιακούς σταθμούς βάσης ευρείας εμβέλειας, -όπως τους γνωρίσαμε και στις προηγούμενες γενιές δικτύων 3G, 4G, κτλ.-, νέοι σταθμοί βάσης, μικρότερης ισχύος εκπομπής –και μικρότερης εμβέλειας- (pico cells, micro cells, femto cells, κτλ.) έχουν ήδη ξεκινήσει να εγκαθίστανται για την βελτιστοποίηση του σχεδιασμού και της εξυπηρέτησης περιοχών με απαιτήσεις οι οποίες δεν μπορούν να αντιμετωπιστούν με τα σημερινά δεδομένα. Σε συνδυασμό με τις μικροκυψέλες τύπου pico, femto, κτλ., στα UDNs συνυπάρχουν και αντίστοιχοι τοπικοί σταθμοί βάσης άλλων τεχνολογιών, όπως WiFi APs. Οι προκλήσεις που δημιουργούνται επομένως, σχετίζονται με το ποιες υπηρεσίες και ποιες κινητές συσκευές θα εξυπηρετηθούν από ποια τεχνολογία/κυψέλη με βέλτιστο τρόπο, έτσι ώστε να ικανοποιούνται οι απαιτήσεις χρήσεις όλων.

Στο δεύτερο κεφάλαιο, αρχικά παρατίθενται κάποιοι βασικοί ορισμοί όρων που χρησιμοποιούνται επανειλλημένα κατά τη διάρκεια της διατριβής, όπως ο ορισμός της επίγνωσης πλαισίου (Context Awareness), των ετερογενών δικτύων (Heterogeneous Networks), η διαχείριση ασύρματων πόρων πρόσβασης (Radio Resource Management), η δρομολόγηση και ο διαχωρισμός της κίνησης δεδομένων (Access Traffic

Steering/Splitting/Switching), η δικτύωση βασισμένη στο λογισμικό (Software Defined Networking), καθώς και η τεχνολογία τεμαχισμού δικτύου (Network Slicing). Στη συνέχεια του 2ου κεφαλαίου, αναλύεται το θέμα των ετερογενών δικτύων. Προβλέπεται ότι η συγκεκριμένη τάση θα οδηγήσει αναπόφευκτα σε πολύ πυκνές υποδομές, στις οποίες, αφενός, οι σταθμοί βάσης LTE θα συνυπάρχουν με την εξέλιξη της 5ης γενιάς τους, ενώ επιπλέον, τα δίκτυα 3GPP συνυπάρχουν με τα μη 3GPP (κυρίως Wi-Fi), δημιουργώντας έτσι μια πολυεπίπεδη αρχιτεκτονική που αποτελείται από ετερογενείς τεχνολογίες ασύρματης πρόσβασης. Μερικές από τις μεγαλύτερες προκλήσεις σε τέτοια πυκνά ασύρματα περιβάλλοντα είναι η αποτελεσματική αλληλεπίδραση και συνεργασία μεταξύ των κυψελωτών συστημάτων διαφορετικών γενεών (3G, 4G, 5G, etc.) και Wi-Fi APs, η βελτιστοποίηση του διαμοιρασμού ασύρματων πόρων και κινητών συσκευών (UE – RAT mapping), καθώς και η ελαχιστοποίηση των περιττών handovers, -όπως και γεγονότων που σχετίζονται με το φαινόμενο ping-pong- μεταξύ γειτονικών κυψελών, οι οποίες αναπόφευκτα επιδεινώνουν την παρεχόμενη ποιότητα υπηρεσιών προς στους χρήστες. Έπειτα, αναλύονται νέες κατευθύνσεις από τις ομάδες προδιαγραφών της 3GPP προς την κατεύθυνση της «σφικτής» διασύνδεσης τεχνολογιών πρόσβασης 3GPP και μη 3GPP (3GPP Tight Interworking), για τη βελτιστοποίηση του προβλήματος της χωρητικότητας του δικτύου. Παρουσιάζονται και αναλύονται τεχνολογίες όπως το Hotspot 2.0, το LWA (LTE-WLAN integration), όπως και νέα πρωτόκολλα για ευέλικτη δρομολόγηση της κίνησης μεταξύ αυτών των τεχνολογιών (π.χ., Local IP Access – LIPTA, Selected IP traffic Offload – SIPTO, Multi-Access PDN Connectivity - MAPCON, IP Flow Mobility – IFOM). Το δεύτερο κεφάλαιο συνεχίζει με την ανάλυση του όρου της επίγνωσης κατάστασης καθώς και την παρουσίαση της σχετικής βιβλιογραφίας.

Στο τρίτο κεφάλαιο παρουσιάζονται οι τρεις βασικοί μηχανισμοί διαχείρισης πόρων (Radio Resource Management): το cell selection/reselection, το call admission control και το handover. Αυτοί οι τρεις μηχανισμοί είναι εκείνοι που επηρεάζουν τελικώς την χαρτογράφηση και τοποθέτηση των κινητών συσκευών στις αντίστοιχες τεχνολογιές πρόσβασης –καθώς και τα αντίστοιχα στρώματα (macro, pico, femto layers, κτλ.). Η διαδικασία του cell (re)selection πραγματοποιείται από την κινητή συσκευή (UE), σε αντίθεση με τις άλλες δύο (call admission control και handover) οι οποίες πραγματοποιούνται από τη μεριά του δικτύου. Σε όλες τις περιπτώσεις, η πρόκληση έγκειται στον εντοπισμό και την πραγματοποίηση της ανάθεσης της βέλτιστης κυψέλης για την εξυπηρέτηση συγκεκριμένων απαιτήσεων μιας κινητής συσκευής από το δίκτυο. Στο τελευταίο κομμάτι του 3ου κεφαλαίου παρατίθενται αναλυτικά όλες οι πρόσφατες βιβλιογραφικές αναφορές που σχετίζονται με τα συγκεκριμένα ερευνητικά κομμάτια.

Στο τέταρτο κεφάλαιο αναλύεται σε βάθος ο μηχανισμός του COmpAsS. Όπως προαναφέρθηκε, το COmpAsS αποτελεί έναν μηχανισμό επιλογής Τεχνολογίας Ασύρματης Πρόσβασης πολλαπλών κριτηρίων με επίγνωση κατάστασης περιβάλλοντας που κάνει χρήση Ασαφούς Λογικής (Fuzzy Logic). Το κύριο μέρος του COmpAsS λειτουργεί στην πλευρά του Εξοπλισμού Χρήστη (UE), ελαχιστοποιώντας τις επιβαρύνσεις σηματοδότησης πάνω από τη διεπαφή αέρα και το φορτίο υπολογισμού στους σταθμούς βάσης. Ο μηχανισμός COmpAsS εκτελεί καταγραφή των συνθηκών του δικτύου σε πραγματικό χρόνο και, σε συνδυασμό με ένα σύνολο προκαθορισμένων κανόνων, υπολογίζει μια λίστα προτεραιότητας που απαρτίζεται από τις καταλληλότερες/διαθέσιμες επιλογές/τεχνολογίες πρόσβασης δικτύου, για κάθε μία από τις ενεργές ροές δεδομένων του UE. Οι συνθήκες δικτύου, οι οποίες καταγράφονται από το COmpAsS και αναλύονται για την δημιουργία της λίστας προτεραιότητας είναι το φόρτο του σταθμού βάσης (αριθμός συνδεδεμένων κινητών συσκευών και ποσοστό χρήσης του συνολικού διαθέσιμου ρυθμού μετάδοσης), το φόρτο του κύριου μέρους του δικτύου στο οποίο συνδέεονται οι σταθμοί βάσης (backhaul link), η μετρική RSRQ (Reference Signal Received Quality) και η οποία σχετίζεται με την ποιότητα του λαμβανόμενου σήματος του

UE από το σταθμό βάσης, η κινητικότητα του UE (mobility), καθώς και το είδος της ροής δεδομένων/υπηρεσίας η οποία μελετάται και συγκεκριμένα η αντίστοιχη ευαισθησία της ροής στην καθυστέρηση (εφόσον αναφέρθηκε ότι η λίστα προτεραιότητας δημιουργείται ανά ροή δεδομένων, και όχι γενικά ανά UE). Το σύστημα Ασαφούς Λογικής που εφαρμόζεται στον συγκεκριμένο μηχανισμό λαμβάνει ως εισόδους τις παραπάνω πέντε παραμέτρους και εξάγει την μετρική της Καταλληλότητας («Suitability»). Η Καταλληλότητα υπολογίζεται για κάθε μία από τις ενεργές ροές δεδομένων (data flows) του UE, είτε αυτές αντιστοιχούν σε διαφορετικές υπηρεσίες/εφαρμογές, είτε στην ίδια. Γενικά, η σχέση που έχει κάθε μια από τις 5 μεταβλητές εισόδου είναι η εξής: όσο μεγαλύτερο είναι το φόρτο ενός σταθμού βάσης, το λαμβανόμενο RSRQ, ή το φόρτο του backhaul link του συγκεκριμένου σταθμού, τόσο λιγότερο κατάλληλη επιλογή θεωρείται. Όσο υψηλότερη είναι η κινητικότητα του UE τόσο καταλληλότερες γίνονται επιλογές μεγάλων κυψελών, όπου και αποφεύγονται με αυτό τον τρόπο οι πολλαπλές και συνεχείς αλλαγές σταθμού εξυπηρέτησης (handovers, ping-pong effects). Τέλος, όσο υψηλότερη είναι η ευαισθησία στην καθυστέρηση δικτύου, τόσο προτιμώνται επιλογές σταθμών με εγγυημένη ποιότητα υπηρεσίας (QoS), όπως για παράδειγμα ένα WiFi AP. Στο 4^{ο} κεφάλαιο, ο παρουσιάζονται όλες οι λεπτομέρειες του αλγορίθμου που εφαρμόστηκε, η αρχιτεκτονική του συστήματος ενταγμένη στο ευρύτερο σύστημα κινητής, η μοντελοποίηση με τη χρήση της Ασαφούς Λογικής, μια αναλυτική προσέγγιση που σχετίζεται με το κόστος σηματοδοσίας του μηχανισμού, καθώς και λεπτομερή αποτελέσματα που παρήχθησαν με τη χρήση του προσομοιωτή NS-3, σε αρκετά σύνθετα περιβάλλοντα που προσομοιάζουν ετερογενή δίκτυα 5^{ης} γενιάς.

Στο πέμπτο κεφάλαιο, παρουσιάζεται ο μηχανισμός του CEPE, μια μηχανή ανάλυσης πληροφορίας πλαισίου και εξόρυξης προφίλ χρηστών (Context Extraction and Profiling Engine). Το CEPE βασίζεται σε μηχανισμούς μηχανικής μάθησης για να αναλύσει την πληροφορία πλαισίου και περιβάλλοντος, τα πρότυπα συμπεριφοράς των χρηστών, να εξάγει σημαντικές γνώσεις και έπειτα εφαρμόσει τις μεθόδους διαχείρισης των πόρων του δικτύου βάσει αυτών των προτύπων συμπεριφοράς/προφίλ για βέλτιστο προγραμματισμό. Στην αρχή του κεφαλαίου παρουσιάζονται συνοπτικά οι προσεγγίσεις της μηχανικής μάθησης που μπορούν να εφαρμοστούν και έπειτα εξηγείται η προσέγγιση που τελικά επιλέχτηκε και ακολουθήθηκε. Το CEPE συλλέγει πληροφορίες σχετικά με τους χρήστες, τις υπηρεσίες που χρησιμοποιούν, τις κινητές συσκευές, καθώς και τις συνθήκες του δικτύου και βασίζεται σε μέθοδο επεξεργασίας χωρίς σύνδεση (offline mode) για να αποκτήσει ένα μοντέλο γνώσης, το οποίο στη συνέχεια χρησιμοποιείται για τη βελτιστοποίηση των κύριων μηχανισμών RRM, δηλαδή το call admission control, το handover και το cell (re)selection. Παρουσιάζεται αναλυτικά το μοντέλο δεδομένων (data model) που εφαρμόζεται, καθώς και οι επιμέρους παράμετροι της κάθε οντότητας που συγκεντρώνει και αναλύει το μοντέλο του CEPE. Επιπλέον, ο μηχανισμός εντάσσεται στο ευρύτερο EPC (Evolved Packet Core) δίκτυο που χρησιμοποιείται στα δίκτυα 4^{ης} γενιάς και παρουσιάζεται η συμβατότητα και οι διεπαφές του με τις υπάρχουσες οντότητες που σχετίζονται με αυτόν. Η βιωσιμότητα και η εγκυρότητα του μηχανισμού επιδεικνύεται μέσω δύο εκτεταμένων σεναρίων προσομοίωσης σε ρεαλιστικά περιβάλλονα 5^{ης} γενιάς με μεγάλο αριθμό τερματικών που κινούνται στοχαστικά και καταναλώνουν διαφορετικές, απαιτητικές –σε ρυθμό μετάδοσης δεδομένων- υπηρεσίες.

Το έκτο κεφάλαιο παρουσιάζει το μηχανισμό CIP (Context Information Preprocessing). Ο συγκεκριμένος μηχανισμός λειτουργεί συμπληρωματικά σε μηχανισμούς ανάλυσης πλαισίου και εξόρυξης πληροφορίας –όπως ο CEPE-, μιας και βασικό αντικείμενό του είναι η ελαχιστοποίηση της πληροφορίας που μεταδίδεται μέσα στο δίκτυο για την επίγνωση πλαισίου, και ως εκ τούτου, η ελαχιστοποίηση του κόστους σηματοδοσίας. Το κύριο σκεπτικό στην υλοποίηση του μηχανισμού είναι το φιλτράρισμα κάθε είδους πληροφορίας που συλλέγεται πριν αποθηκευτεί για να αποσταλεί στη μηχανή εξόρυξης

πληροφορίας (π.χ. CEPE από εδώ και στο εξής). Κάθε δικτυακό στοιχείο διαθέτει την μοντελοποιημένη πληροφορία (π.χ. προφίλ χρηστών) που έχει εξάγει το CEPE μέσω μιας κατανεμημένης αρχιτεκτονικής. Σε τακτά διαστήματα, τα προφίλ αυτά ανανεώνονται σε όλα τα δικτυακά στοιχεία, ούτε ώστε όλα να είναι ενημερωμένα με την πιο πρόσφατη βάση προφίλ χρηστών. Κάθε φορά που ένα δικτυακό στοιχείο καταγράφει κάποιο δεδομένο που πιθανώς «ενδιαφέρει» το CEPE για να ανανεώσει τη βάση των προφίλ του, πραγματοποιεί έναν συσχετισμό με το αποθηκευμένο προφίλ του συγκεκριμένου UE, το οποίο αφορά η πληροφορία. Εάν το νεοαποκτηθέν δεδομένο είναι σύμφωνο με το προφίλ που ήδη έχει παράξει το CEPE, τότε θεωρείται περιττό προς εκπομπή, οπότε και απορρίπτεται, ενώ ενημερώνεται ένας σχετικός δείκτης-μετρητής (Consistency Index). Ένα αναλυτικό παράδειγμα δίνεται στο οποίο βήμα προς βήμα η διαδικασία αποθήκευσης ή απόρριψης πληροφοριών πλαισίου. Επιπλέον, στο τελευταίο μέρος του κεφαλαίου παρατίθεται μια αναλυτική μαθηματική προσέγγιση η οποία ποσοτικοποιεί το κέρδος χρήσης του προτεινόμενου μηχανισμού –όσον αφορά το κόστος σηματοδοσίας- σε ρεαλιστικά δικτυακά περιβάλλοντα 5^{ης} γενιάς.

Στο έβδομο κεφάλαιο, παρουσιάζεται το προτεινόμενο πλαίσιο μηχανισμών (CEPE, COmpAsS και CIP) σαν ένας ενιαίος μηχανισμός και εντάσσεται στην τελευταία, προτεινόμενη αρχιτεκτονική της 3GPP για τα δίκτυα 5^{ης} γενιάς, συνδέοντας την προτεινόμενη λειτουργικότητα των 3 μηχανισμών με τις νέες δικτυακές οντότητες που περιγράφονται στο νέο σύστημα 5G. Αρχικά, παρατίθεται μια συζήτηση η οποία αναλύει τη συμπληρωματικότητα των δύο βασικών μηχανισμών διαχείρισης πόρων – CEPE και COmpAsS-, βάσει της βασικής διαφοράς στη λειτουργικότητά τους: το CEPE λειτουργεί από τη μεριά του Core δικτύου, ενώ το COmpAsS από τη μεριά του UE. Όπως είναι αναμενόμενο, συγκεκριμένα πλεονεκτήματα και μειονεκτήματα της κάθε προσέγγισης –όταν αυτές λειτουργούν ξεχωριστά-, αποδεικνύουν τη χρησιμότητα και τη βελτιστοποίηση στη διαχείριση των πόρων του δικτύου κατά την παράλληλη και συντονισμένη λειτουργία τους: Οι βασικές πολιτικές του δικτύου δίνονται από το CEPE. Αυτές είναι εκείνες όπου θα καθορίσουν τελικά τους πόρους που θα διατεθούν για μια συγκεκριμένη ροή δεδομένων ενός UE. Κάθε UE από την άλλη, έχει τη δυνατότητα με τη χρήση του COmpAsS να πληροφορεί το δίκτυο (και ως εκ τούτου, το CEPE), για τη βέλτιστη επιλογή –σε πραγματικό χρόνο- μιας ήδη ενεργοποιημένης συνεδρίας. Το CEPE έχει τη δυνατότητα σχεδιασμού για μια συνολικότερη δικτυακή περιοχή, χρησιμοποιώντας και της δυνατότητες πρόβλεψης χρήσης υπηρεσιών και κινητικότητας των UE που διαθέτει. Σε σχέση με την αρχιτεκτονική του 5G δικτύου, παρουσιάζονται δύο νέα δικτυακά στοιχεία/οντότητες, τα οποία περιγράφονται προσφάτως στις τελευταίες δημοσιεύσεις της 3GPP: πρόκειται για το Network Data Analytics Function (NWDAF) και το Access Traffic Steering, Switching and Splitting (ATSSS). Το CEPE, ως αναλυτής δεδομένων και μηχανισμός πρόβλεψης δεδομένων, είναι ουσιαστικά ένα δομοστοιχείο ικανό να υποστηρίξει πλήρως αυτή τη λειτουργία του NWDAF, αποτελώντας μια συγκεκριμένη υλοποίησή του, στην επικείμενη αρχιτεκτονική 5G. Το NWDAF-CEPE θα παρέχει πληροφορίες σε σχέση με το φόρτο σε επίπεδο “τεμαχίου” δικτύου (network slice). Με τη χρήση της τεχνολογίας της Δικτύωσης Βασισμένης στο Λογισμικό (SDN), δίνεται η δυνατότητα δημιουργίας πολιτικών και κανόνων που δημιουργούν υποδίκτυα/τεμάχια με συγκεκριμένα χαρακτηριστικά και τα οποία ανατίθενται δυναμικά στις υπάρχουσες κινητές συσκευές. Επιπλέον, μέσω συγκεκριμένων διεπαφών που αναπτύσσονται μεταξύ του NWDAF και της οντότητας που θα διαχειρίζεται τους μηχανισμούς του slicing στο 5G (όπως το Network Slice Selection Function - NSSF), όπως και διεπαφών μεταξύ του NWDAF και του PCF (Policy Control Function, όπως και στο EPC), το NWDAF-CEPE θα μεταφέρει προφίλ, πρότυπα πρόβλεψης και πολιτικές στις οντότητες αυτές. Το ATSSS είναι υπεύθυνο για τη διαχείριση των ροών δεδομένων των UE και την σύνδεση της κάθε ροής με την αντίστοιχη βέλτιστη, διαθέσιμη τεχνολογία πρόσβασης. Η αντιστοίχιση αυτή θα βασίζεται πρωτίστως στις πολιτικές που λαμβάνονται από το NWDAF-CEPE, αλλά

δευτερευόντως και από το UE βάσει παραμέτρων πλαισίου που θα οριστούν. Με αυτόν τον τρόπο δείχνεται πώς COmpAsS και CEPE συνεργάζονται στο πλαίσιο της νέας αρχιτεκτονικής 5ης γενιάς. Τέλος, το CIP θα προεπεξεργάζεται την πληροφορία που ανταλάσσεται μέσω όλων των προαναφερθείσων διεπαφών, ενώ το COmpAsS θα εκφράζει από τη μεριά του 5G-UE τις βέλτιστες επιλογές σε πραγματικό χρόνο, οι οποίες θα καταγράφονται και θα αξιολογούνται από το δίκτυο (NWDAF-CEPE).

Στο προτελευταίο, όγδοο κεφάλαιο, παρουσιάζονται κάποιες συμπληρωματικές συνεισφορές της διατριβής. Το πρώτο κομμάτι παρουσιάζει μια μελέτη σχετική με Traffic Engineering, πρόβλεψη φόρτου και αριθμού χρηστών, με αντίστοιχη διαχείριση και μελέτη του δικτύου, η οποία έχει βασιστεί ολοκληρωτικά στο CEPE. Το δεύτερο κομμάτι παρουσιάζει μια συγκεκριμένη εφαρμογή Smart Farming στα πλαίσια του Internet of Things, και των απαιτήσεων που αυτό παρουσιάζει για το σύστημα κινητών δικτύων 5ης γενιάς.

Τέλος, στο ένατο κεφάλαιο συνοψίζεται η ερευνητική συνεισφορά της διατριβής. Επιπλέον αποτυπώνονται τα προβλήματα που εντοπίστηκαν και οι προτεινόμενες λύσεις. Η τρέχουσα διατριβή καταλήγει παρουσιάζοντας τις σημαντικότερες προεκτάσεις των ερευνητικών θεμάτων που μελετήθηκαν κατά την εκπόνηση της διατριβής και ενδιαφέρουσες μελλοντικές επεκτάσεις.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Towards 5G Networks

In the following years, the number of wireless and mobile devices is expected to increase considerably. Along with it, a huge increase of mobile traffic [1] will also take place. More specifically, the mobile traffic in 2016 was nearly 30 times the size of the entire global Internet in 2000. Almost half a billion mobile devices and connections were added in 2016, while at the same time, average smartphone usage grew 20% in the same year. In addition, the actual traffic volume per subscriber increases 25-40% per year, thus exceeding the expectations set by ITU [2]. The deployment of 5G cellular networks targets to support this vast number of devices, while at the same time existing 3GPP specifications will keep on supporting legacy cellular access networks (e.g., GSM, HSPA, LTE, LTE-A), as well as alternative radio access technologies (e.g., Wi-Fi). In this environment, the end users will have access to a diverse set of services (high definition video and audio, web browsing, games, etc.). It is worth pointing out in parallel that the high penetration of smartphones and tablets on the market [3] will enable end users to make use of all these services while on the move.

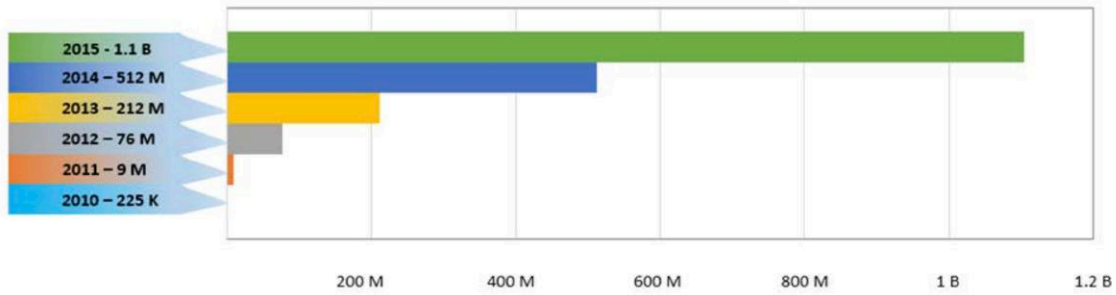


Figure 1: LTE Connections 2010 – 2015 [1]

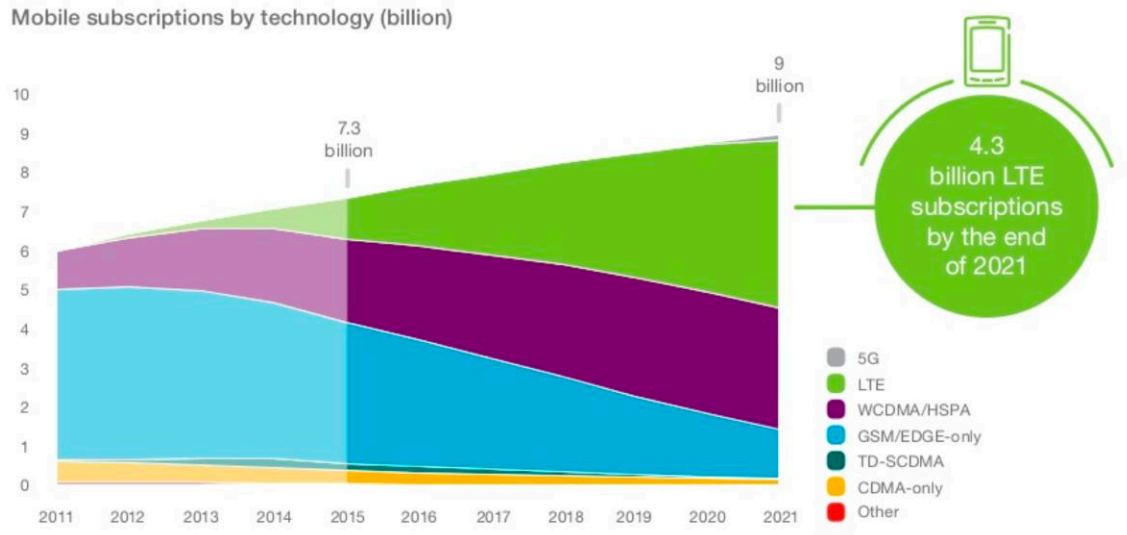


Figure 2: LTE and 5G Forecast 2016 – 2021 [2]

5G networks are expected to support billions of small end devices (e.g., sensors,

[1] World Cellular Information Service (WCIS), Ovum, October 2016.

[2] Ericsson mobility report 2016

actuators, etc.) as well as communicating vehicles [4] in the context of Machine Type Communication (MTC). The vision is that 5G networks will manage to materialize the Internet of Things (IoT) ecosystem. This realization will unveil new requirements to the network operators and telecom manufacturers. Significant problems arise due to the fact that these -billions of- devices, which communicate short messages in periodic or asynchronous mode, will compete with typical User Equipment (UE) devices for the 5G resources. Moreover, these devices will support a variety of services with their own requirements (e.g., short packet size, burst traffic, sensitivity to delay or losses). Besides the tremendous growth, which is expected in terms of number of devices, due to an increasingly diverse set of new and yet unforeseen services, users and applications (including machine-to-machine (M2M) modules, Smart Cities, industrial automation, etc.), novel and less predictable mobile traffic patterns are also expected to emerge. The aforementioned requirements should also be taken into consideration by the underlying network infrastructure. Overall, existing mechanisms used for the communication of end terminals are inadequate to support the future needs. Towards improving efficiency, well-established mechanisms have to be redesigned. One of the most important areas 5G networks have to improve is the mapping of smart devices and services to different RATs and layers (i.e., macro, micro, pico, femto cells). This mapping affects the Key Performance Indicators (KPIs) of a network in relation to the experienced grade of service (e.g., blocking probabilities, throughput, delay, jitter, etc.). The placement of a UE to a RAT or a cell layer, that is either in idle or a connected mode, is primarily realized via three vital RRM mechanisms, namely: a) Cell-(re)Selection (CS), b) Access/Admission Control (AC) and c) Handover (HO).

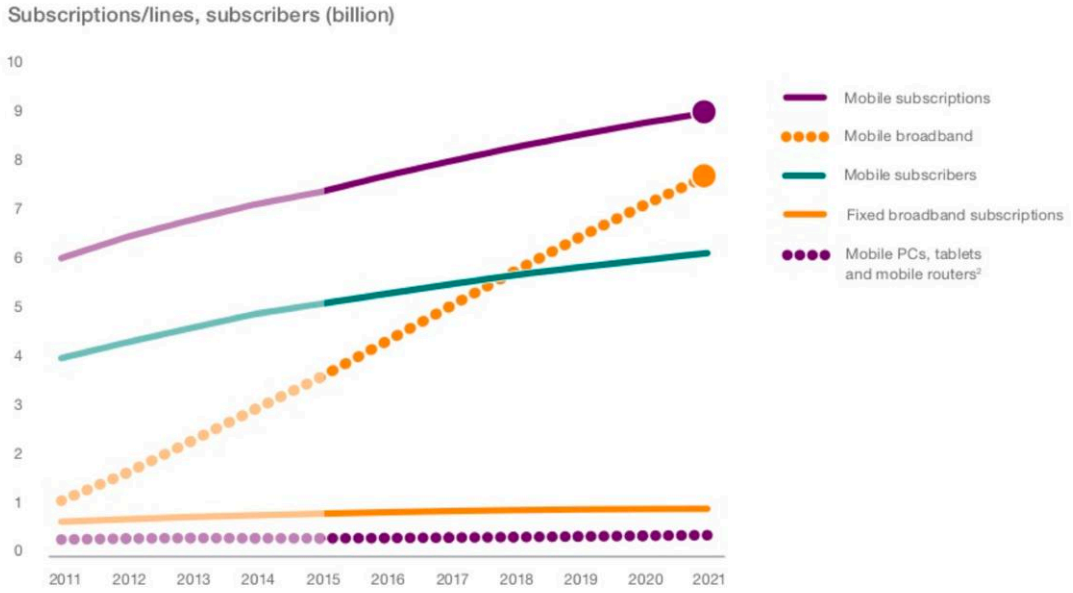


Figure 3: Growth of mobile subscriptions[3]

1.2 Motivation and Research Challenges

Network densification manifesting in deployments of Small Cells (SCs) is an ongoing trend in contemporary cellular networks. Although SCs were already commercially available for the 2G and 3G technologies, the LTE and LTE-A standards provide technical solutions that exploit the local nature of such deployments. SCs are well suited for handling large traffic demands in hotspot areas with noticeable proliferation over the last years of high-end devices capable of processing data heavy content, e.g. high definition video. Moreover, people expect to have a broadband experience not only at home or

[3] Ericsson mobility report 2016

office, but also outdoors. These two trends combined create a massive upsurge of cellular traffic, often referred to as the x1000 traffic volume challenge [5]. The next generation of cellular technology –5G– is expected to provide an economically justified system that will cater for this massive demand and extravagant user requirements.

The performance of modern cellular networks, mainly limited by the radio access network, is usually enhanced through solutions aiming at improving spectral efficiency, such as advanced antenna techniques (including the use of massive number of antennas) and endeavors of cellular industry to obtain more spectrum for wireless transmission in low and high frequency bands [5]. Despite technical challenges, this way forward is definitely a promising direction to improve capacity of future 5G networks, but, without a doubt, they will not be sufficient to provide a ubiquitous high-end user experience for the 2020-and-beyond mobile society. As proven in contemporary cellular networks, in order to satisfy growing user demands, improved spectral efficiency should be accompanied by further cell densification, especially in dense urban areas and indoors. Massive roll-out of SCs immediately poses a question on its economic feasibility. SC solutions available today rely on methods such as distributed antenna systems, unlicensed spectrum, or user-deployed SCs in order to bring down the deployment costs. SCs can be also extended to moving relays or nomadic cells where antenna systems exploiting wireless backhaul are mounted on cars, buses or trains, in order to provide a broadband experience to users inside or in proximity of vehicles.

The above-mentioned factors suggest that further deployment densification, resulting in UDNs is inevitable, which has interesting consequences for future networks operations. Shrunken cell sizes lead to reduced number of users served simultaneously by individual SCs over a geographical area, and hence to sharing the radio resources among fewer users. Moreover, smaller user-to-access-node distances decrease the probability of severe shadowing. This factor plays a major role in wave propagation at higher frequencies, which are interesting due to the availability of large bandwidths. **Higher frequencies are a perfect fit for UDNs since, paradoxically, their higher attenuation limits the interference to neighbouring sites and users**. On the other hand, **fewer users per cell leads to a more bursty activity profile of SCs**. In combination with the time division duplexing (TDD) mode, which is expected to be extensively used in 5G due to its capability to adapt to dynamic traffic demands, this will pose a significant challenge to future 5G resource allocation schemes. It is still an open question to which extent advanced receivers and transmission schemes will be able to cope with the dynamic interferences [6].

Another challenge expected in UDNs is the heterogeneity of the 5G deployment, i.e. the 5G UDN Heterogeneous Networks (HetNets). 5G is expected not only to introduce new access technologies, but also to reuse legacy 3GPP systems as well as IEEE technologies in order to provide the required user experience exactly where it is needed. This complicated deployment is very demanding from the mobility point of view, but it is also an opportunity for future devices to use specific technologies or layers in order to provide the necessary performance. How to efficiently detect and exploit this heterogeneous environment is definitely one of the most important challenges for the UDN design.

All aforementioned factors pose a question mark on the interference, resource and mobility management schemes that are used in current cellular networks, and they call for new methods, which will be able to fully exploit the benefits of heterogeneous UDN deployments of the 5th generation networks.

1.3 Thesis Contribution

The contributions of this dissertation move towards the following major directions:

a) COmpAsS, a user-oriented context-based scheme for RAT selection and traffic steering/switching, which processes context in real-time and produces a RAT suitability list to be used for handover management reasons,

b) CEPE, a knowledge extraction engine based on data mining techniques towards user profiling and network policies formulation,

c) CIP, a context information pre-processing scheme, which acts in an augmenting manner to the former two, in order to minimize the high context acquisition signaling overhead.

d) A study on CEPE-COmpAsS interworking in an SDN-enabled, 5G architecture, capable of applying network slicing approaches.

Supplementary contributions, which enforce the research carried out in the aforementioned primary four directions and also comprise parts of the next steps to be made in the context of this research domain are:

- the architectural perspective of the proposed schemes, which takes into account the latest 3GPP standardization guidelines and attempts to provide a valid and viable solution towards the forthcoming 5G architecture,
- a study on network traffic engineering policies, which can exploit CEPE and user profiling methodology is included,
- an attempt to describe from a common point of view 3 primary RRM mechanisms, i.e. cell (re)selection, handover and call admission control via a comprehensive categorization of the existing approaches, both from the academic area, as well as from the industry, by incorporating the available patents as well, and finally
- a 5G use case application related to IoT and Precision Farming, which highlights specific requirements related to industrial applications, ultra-low delay requirements, etc.

The first major contribution of this thesis is **COmpAsS**, a context-aware, user-oriented RAT selection mechanism, which operates on the User Equipment (UE) side and ultimately produces a list of the most suitable RATs per active traffic flow/session, towards QoS optimization. One of the greatest advantages of the UE-based solution is the minimization signaling overhead over the air interface, as well as the computation load on the base stations. COmpAsS collects information related to the network status, such as the load of the base stations, the load of the backhaul link, the Reference Received Signal Quality (RSRQ), user mobility information, such as the velocity of the UE, as well as the specific QoS requirements of the type of traffic to be transmitted, in order to assess -in real-time- the most suitable RAT and/or cell layer, which should serve the UE's active sessions. COmpAsS mechanism adopts Fuzzy Logic (FL) as one the core logic modules, responsible for the perception of the network situation and, in combination with a set of pre-defined rules, calculates a list of the most suitable available access network options. Furthermore, we propose an evolution of 3GPP's Access Network Discovery and Selection (ANDSF) function, as one of the primary Evolved Packet Core (EPC) network functions collaborating with COmpAsS for the exchange of the required context information. The merits of COmpAsS are showcased via an extensive series of simulation scenarios, as part of 5G ultra dense networks (UDN) use cases. The results prove how the proposed mechanism optimises Key Performance Indicators (KPIs), when juxtaposed to a well-established LTE handover algorithm.

The second major contribution of the current thesis **the Context Extraction and Profiling Engine (CEPE)**, a resource management framework, which collects diverse types of context information and performs data mining techniques in order to extract

meaningful knowledge. The context information, which is aggregated, primarily relates to four categories: **network operation data, user behavior information, terminal capabilities and application/service data**. CEPE analyzes this information, extracts meaningful knowledge and performs user profiling in order to apply it for optimal resource planning, as well as prediction of resource requirements. CEPE collects information about users, services, terminals and network conditions and –based on offline processing– derives a knowledge model, which is subsequently used for the optimization of the primary RRM mechanisms, i.e. handover, cell selection and call admission control. From a methodological point of view, initially the KPIs that will be employed are identified in order to assess the efficiency of the mechanism. Next, the types of data that should be monitored are identified (network operation data, user behaviour information, etc.). Then, the extracted context information is translated into user profiles and is finally applied as input for enhanced cell (re)selection, handover or admission control. CEPE's operation is tightly connected to the scheme, which follows, CIP, and focuses on the pre-processing of the vast amount of information, which is collected, towards minimizing the signaling overhead. The viability and validity of CEPE is demonstrated via an extensive set of simulation scenarios.

The third major contribution is **CIP**, a Context Information Pre-processing scheme, aiming to identify and discard redundant or unnecessary data before knowledge extraction. CIP could be considered as an integral part of the afore described profiling schemes, i.e. COmpAsS and CEPE. CIP comprises a framework that primarily relies upon data aggregation and pre-processing techniques. Context information processing and knowledge extraction is considered a great tool towards the optimisation of several network functions; nevertheless, the acquisition of the context is often a very costly process –in terms of signaling burden imposed on the network. The module comprises aggregating and compressing mobile network-related context information per unique identifier, such as the end device's International Mobile Subscriber Identity (IMSI), as well as techniques related to identifying and discarding user profile-redundant or unnecessary context data, before any transmission to CEPE. CIP gains are illustrated via a detailed analytical approach, guided by well-established 5G use case requirements.

The fourth major contribution of this thesis is a mapping of the proposed scheme in a **Software Defined Networking-enabled 5G architecture**, as proposed by the latest 3GPP standardization, capable of applying Network Slicing approaches for further optimizing the network resources distribution and sharing and addressing the challenging 5G use cases, such as massive IoT.

1.4 The Structure of the Thesis

The current thesis is structured into nine chapters. Following the current chapter, the structure is briefly presented below:

Section 2 provides the background of the specific domain and introduces the main concept to the reader. The definitions of the primary concepts and terminology is provided. Next, a sub-section focusing on the Heterogeneous Radio Access Technologies towards 5G is provided, which describes in detail the challenges, use cases and characteristics of the forthcoming 5G environment. At the end of Section 2, the notion of Context Awareness is defined, while a discussion takes place as well with regard to the Data Analytics approaches for Network Resource Management.

Section 3 provides a comprehensive review of the literature and industry patents, related to Radio Access Technology (RAT) selection processes, and more specifically, the three primary operations related to RAT selection, i.e., Cell (Re-)selection, Admission Control and Handover.

Section 4 describes COmpAsS, the first of the three primary contributions of this thesis. An overview of the proposed solution is provided, along with a discussion with regard to the architectural aspect. A detailed description of COmpAsS algorithm is given, as well as a comprehensive definition of the Fuzzy Logic-based component. The signaling costs of the proposed scheme are quantified, targeting to evaluate the validity of COmpAsS. The last part of the section, focuses on the experimental evaluation of the scheme.

Section 5 focuses on the second proposed mechanism, CEPE. Similarly with Section 4, an overview of the mechanism is initially provided. Next, CEPE deployment in the EPC network architecture is discussed in detail, while in the last part of Section 5, the evaluation of the proposed framework takes place, via an extensive experimental evaluation.

Section 6 provides a comprehensive description of the third primary contribution of the thesis, i.e. CIP. An overview of the Context Information Pre-Processing engine is initially provided. Later in the section, an analytical approach is followed in order to evaluate CIP and quantify the overall system signaling reduction gains.

Section 7 make a link between the three proposed mechanisms, i.e., COmpAsS, CEPE and CIP, in the context of the latest 5G proposed architecture, as described in the latest standardization by 3GPP. A detailed description of the mechanisms' roles in the novel system is provided, along with a mapping between the novel 5G network components, and the proposed schemes.

Section 8 provides insights in relation to the supplementary work, which took place in the context of the thesis, and is not indirectly link to the aforementioned solutions.

Section 9 finally, concludes this work, makes an overall assessment of the thesis contributions and architectural proposals and identifies potential future research directions.

2. BACKGROUND

In this section, the basic principles and notions are being introduced, along with an initial discussion on the specific characteristics and requirements that emerge due to the forthcoming 5G heterogeneous environments. Initially, a detailed discussion on the Heterogeneous Networks takes place, and the directions of the standardization bodies is provided. Technologies such as the Access Network Discovery and Selection Function and Hotspot 2.0 are introduced, which are of upmost importance for the forthcoming developments of 5G HetNets. The LTE-Wi-Fi interworking is highlighted, as one of the highest priorities towards network densification. The initiatives of the aforementioned interworking are provided, as well as the latest efforts from the literature. Both technologies are considered potential core parts of the forthcoming 3GPP – WLAN integration activities. The second part of this section brings in the notion of Context, one of the core concepts of this thesis, on which the overall proposed RRM procedures rely on. Besides the analysis that takes place, a comprehensive listing of the state of the art, which relates to Context is provided.

2.1 Definitions

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

Radio Access Technology or **RAT** is the underlying physical connection method for a radio based communication network. Many modern phones or User Equipment (UEs) in general support several RATs in one device such as Bluetooth, Wi-Fi, and 3G, 4G or LTE. More recently, the term RAT is used in discussions of heterogeneous wireless networks. The term is used when a user device selects between the type of RAT being used to connect to the Internet. This is often performed similar to access point selection in IEEE 802.11 (Wi-Fi) based networks. Tightly linked to the notion of RAT is the Radio Access Network.

Radio Access Network or **RAN** is the part of a mobile telecommunication system, which implements a RAT. Conceptually, it resides between a device such as a mobile phone, a computer, or any remotely controlled machine and provides connection with its core network (CN). Examples of radio network types are GRAN (GSM radio access network), GERAN (similar with GRAN but specifying also the inclusion of EDGE packet radio services), UTRAN (UMTS radio access network), E-UTRAN (the LTE access network).

Radio Resource Management or **RRM** is the system level management of co-channel interference, radio resources, as well as other radio transmission characteristics in wireless communication systems, such as cellular networks, wireless local are networks (WLANs) and wireless sensor systems (WSNs). RRM incorporates policies and algorithms related to power transmission, user allocation, data rates, handover criteria, modulation scheme selection, scheduling, etc. The ultimate objective is to utilize the limited radio-frequency spectrum as efficiently as possible.

Heterogeneous Network or **HetNet** is described the type of network deployment, in which macro-, micro-, pico- and femto- cells from diverse Radio Access Technologies (such as LTE, Wi-Fi, Bluetooth, etc.) operate concurrently. This densification improves network coverage, improves spectral efficiency and reduces the UE power consumption due to the nearby cell. At the same time, however, critical challenges emerge mainly related to intelligent dynamic traffic steering of the network, as well as interference management, due to the uncoordinated manner that small cells usually operate.

Context is defined as the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.

Context Awareness or **CA** in the research area of pervasive computing is firstly employed in [153] and generally refers to the ability of computing systems to acquire and reason about the context information and adapt accordingly the corresponding applications. Context, as introduced by Dey in [157], "is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves".

Access Traffic Steering: The procedure that selects an access network for a *new data flow* and transfers the traffic of this data flow over the selected access network. Access traffic steering is applicable between 3GPP and non-3GPP accesses.

Access Traffic Switching: The procedure that moves all traffic of *an ongoing data* flow from one access network to another access network in a way that maintains the continuity of the data flow. Access traffic switching is applicable between 3GPP and non-3GPP accesses.

Access Traffic Splitting: The procedure that splits the traffic of a data flow across multiple access networks. When traffic splitting is applied to a data flow, some traffic of the data flow is transferred via one access and some other traffic of the same data flow is transferred via another access. Access traffic splitting is applicable between 3GPP and non- 3GPP accesses.

Software Defined Networking: A novel technology, which facilitates network management by enabling programmatically efficient network configuration in order to improve network performance and monitoring. SDN centralizes network intelligence in one network component by disassociating the forwarding process of network packets (Data Plane) from the routing process (Control plane).

Network Slicing: A procedure that enables flexibility, as it allows multiple logical networks to be created on top of a common shared physical infrastructure.

2.2 Heterogeneous Radio Access Technologies Towards 5G

Heterogeneous Networks or HetNets is described the type of network deployment, in which macro-, micro-, pico- and femto- cells (3GPP or non-3GPP such as WiFi) operate concurrently. Due to the high demand of very high data rates in future 5G applications, as well as the ultra-low latency requirements, which are posed, the network operators have already deploying hundreds of small cells in limited geographical areas, i.e. densifying the radio environment, in order to increase the access bandwidth. This densification improves network coverage, improves spectral efficiency and reduces the UE power consumption due to the nearby by cell. At the same time, however, critical challenges emerge mainly related to traffic steering choices of the network, as well as interference management, due to the uncoordinated manner that small cells usually operate.

It is envisaged that the aforementioned trend will inevitably result in very dense deployments, in which on the one hand, UMTS or Long Term Evolution (LTE) base stations (BSs) will co-exist with their 5G evolution, while in addition, 3GPP networks will co-exist with the non-3GPP ones (primarily Wi-Fi), creating thus a multi-tier architecture consisting of heterogeneous radio access technologies. Some of the greatest challenges in such dense wireless environments are the efficient inter-working between the legacy with the latest cellular systems, as well as with Wi-Fi APs, the optimization of the UE placement – RAT selection procedures, as well as the minimization of the unnecessary handovers – and ping-pong effect-related events – between adjacent RATs and cells, which inevitably deteriorate the provided QoS to the users: The handover procedure in the current Evolved Packet Core (EPC)/4G system involves latency overheads, even in

limited coverage areas over the GPRS Tunneling Protocol (GTP) tunnel. In order to enable seamless UE mobility when moving across the different (H)eNBs, the S-GW (at the network core) communicates with the eNBs (at the network edge) to perform handover management; QoS allocation, traffic condition monitoring, user terminal mobility management and security tasks are also forwarded to the Packet Gateway (P-GW). At the same time, the eNB, the S-GW, and the P-GW perform several signaling procedures to handle the session setup at different levels. Such an approach decreases considerably the network performance by increasing the latency and thereby reducing the QoS required for the future real-time applications. Thus, it becomes of utmost importance that frequent or unnecessary handovers in such ultra-dense network environments are minimized; latency overheads should be minimized and the optimal RAT options for the UEs should be available in an efficient way via a viable RAT selection approach.

2.2.1 LTE – WLAN Interworking Standardization activities

Lately, new directions have been presented by 3GPP's specification groups ([7]) towards the network capacity issue optimization and the so called tight interworking of 3GPP and non-3GPP access technologies, with plenty of these novel directions and standards already partially integrated and planned to be fully integrated in the forthcoming releases. In relation to the efficient interworking between heterogeneous wireless systems (e.g., LTE and Wi-Fi), although during the last decade there has been considerable progress in terms of specifications and standards, still a successful demonstration of a seamless integration of Wi-Fi APs with commercial cellular networks, and in realistic 5G business cases is missing. This is because of a number of reasons. Wi-Fi suffers from interference issues since it operates on the unlicensed spectrum. Most importantly however, switching from a cellular network to a Wi-Fi access point has not properly yet evolved to a fully transparent process -from different perspectives-; for the end users, the authentication process had to take place manually – thus, deteriorating the QoE –; further- more, the mobility of multiple flows (even of the same service) among different PDN connections and access technologies was only recently standardized and described [130]. In addition to the first point – and as this is described by Hotspot 2.0 –, all the "islands" of hotspots should be also interconnected into larger "footprints" via further roaming agreements between Wi-Fi operators. Finally, there are still diverse strategies in the way non-3GPP networks are handled by different devices and operating systems, meaning that the software that handles the active UE connections (e.g. the "Connectivity Manager" in Android) has not been standardized. In some cases, even the same OS handles differently the connections de- pending on the version of the OS, e.g. [131].

One of the relevant technologies currently being defined and standardized by 3GPP is the LTE-WLAN integration, i.e. LWA ([40]). LWA has been standardized by the 3GPP in Release-13. Release 14 Enhanced LWA (eLWA) adds support for 60GHz band (802.11ad and 802.11ay aka WiGig) with 2.16 GHz bandwidth, uplink aggregation, mobility improvements and other enhancements.

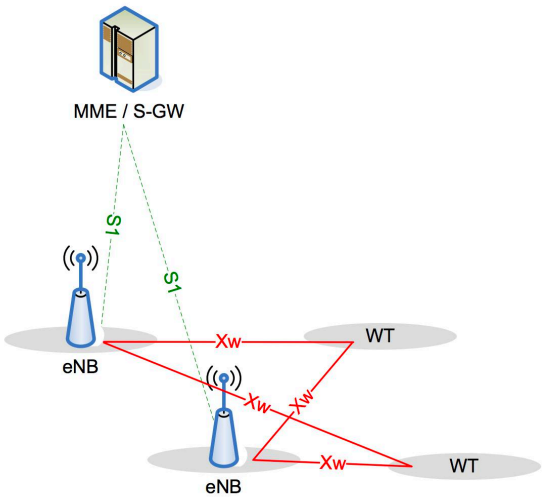


Figure 4: LWA: Network Architecture

In LWA, a mobile handset supporting both LTE and Wi-Fi may be configured by the network to utilize both links simultaneously. It provides an alternative method of using LTE in unlicensed spectrum, which unlike LAA/LTE-U can be deployed without hardware changes to the network infrastructure equipment and mobile devices, while providing similar performance to that of LAA. Unlike other methods of using LTE and WLAN simultaneously (e.g. Multipath TCP), LWA allows using both links for a single traffic flow and is generally more efficient, due to coordination at lower protocol stack layers.

For a user, LWA offers seamless usage of both LTE and Wi-Fi networks and substantially increased performance. For a cellular operator, LWA simplifies Wi-Fi deployment, improves system utilization and reduces network operation and management costs. LWA can be deployed in collocated manner, where the eNB and the Wi-Fi AP or AC are integrated into the same physical device or in non-collocated manner, where the eNB and the Wi-Fi AP or AC are connected via a standardized interface referred to as *Xw*. The latter deployment option is particularly suitable for the case when Wi-Fi needs to cover large areas and/or Wi-Fi services are provided by a 3rd party (e.g. a university campus), rather than a cellular operator.

From the network perspective, there are two options that provide flexibility when looking at deploying LWA - collocated and non-collocated. LWA design primarily follows LTE Dual Connectivity (DC) architecture as defined in 3GPP Release 12, which allows a UE to connect to multiple base stations simultaneously, with WLAN used instead of LTE Secondary eNB (SeNB).

In the user plane, LTE and WLAN are aggregated at the Packet Data Convergence Protocol (PDCP) level. In the downlink, the eNB may schedule PDCP PDUs of the same bearer to be delivered to the UE either via LTE or WLAN. In order to perform efficient scheduling and to assign packets to LTE and WLAN links in the most efficient manner, the eNB can receive radio information about both links, including flow control indication. In order to avoid changes to the WLAN MAC, LWA uses an EtherType allocated for this purpose, so that LWA traffic is transparent to WLAN AP.

In the control plane, Evolved Node B (eNB) is responsible for LWA activation, de-activation and the decision as to which bearers are offloaded to the WLAN. It does so using WLAN measurement information reported by the UE. Once LWA is activated, the eNB configures the UE with a list of WLAN identifiers (referred to as the WLAN Mobility Set) within which the UE can move without notifying the network. This is a tradeoff between fully network controlled mobility and fully UE controlled mobility. Even though

WLAN usage in LWA is controlled by cellular network, UE has the option to "opt out" in order to use home WLAN (in case UE does not support concurrent WLAN operation).

In parallel to LWA, another technology from 3GPP standardization activities, LTE WLAN Radio Level Integration with IPsec Tunnel - LWIP ([40]) is also on its way. LWIP is similar to LWA, in the sense that both make use of unlicensed 802.11 technologies. The difference is that LWA aggregates LTE and Wi-Fi at the PDCP layer, while LWIP aggregates or switches between LTE and Wi-Fi links at the IP layer, just above PDCP. Although both LWA and LWIP have been defined to integrate 802.11-based access networks into the E-UTRAN access network, LWIP has been specifically architected to be able to leverage legacy WLAN infrastructure. Both LWA and LWIP can be contrasted with LAA and its pre-standard version LTE-U that use a modified LTE waveform in the unlicensed band instead of Wi-Fi. Further, in Release 13, LWIP is the only technique defined by 3GPP that is able to send uplink data over the unlicensed spectrum, with the others approaches being purely focusing on enhancing the LTE downlink capabilities.

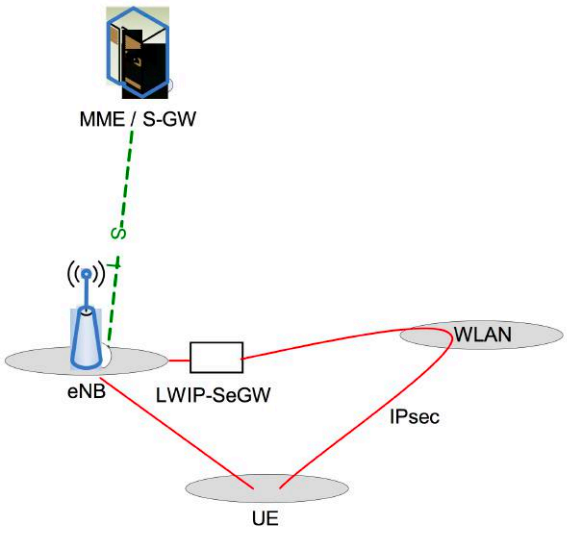


Figure 5: LWIP: Network Architecture

Besides LWA and LWIP, 3GPP has worked towards additional RAT integration solutions that depend upon whether the core network is a UMTS Core network (I-WLAN, [8]) or an Enhanced Packet Core (EPC) network ([9], [10]). The current dissertation focuses on the latter category. The new standards introduce trusted and non- trusted access networks mainly from the point of view of the operator. Thus, they require additional security mechanisms, they provide connections of the Policy and Charging Rules Function (PCRF) to gateway functions and they define the Access Network Discovery and Selection Function (ANDSF, [11]). The ANDSF functionality enables the operator to store policies for the discovery and selection of RATs (e.g. Wi-Fi hotspots) in a server and communicate them to UEs via push or pull methods.

For the interworking between 3GPP and non-3GPP networks, mobility of IP-Flows among them is defined. Mobility maybe handled differently depending if the IP address of UE is preserved (seamless) or not (non-seamless), or if it is handling on per IP-flow or on a per UE basis. Different mobility solutions take into consideration the use of the GPRS Tunneling Protocol (GTP), Proxy Mobile IP (PMIP) or Dual Stack Mobile IP (DSMIP). In all cases, the mobility process is triggered by the UE and not by the network, although efforts are underway to standardize network initiated mobility. Furthermore, as described in [12], a number of protocols (e.g., Local IP Access – LIPTA, Selected IP traffic Offload – SIPTO, Multi-Access PDN Connectivity - MAPCON, IP Flow Mobility - IFOM) are designed for flexible traffic steering.

At the same time, the Wi-Fi Alliance has defined in the context of Hotspot 2.0, the means to assist in the selection of Wi-Fi hotspots and address security issues that today are complex and non-transparent to the users. Hotspot 2.0 standard from Wi-Fi Alliance improves the ability of WLAN devices to discover and connect in a secure way to public Wi-Fi APs. Hotspot 2.0 builds on 802.11u specifications ([132]) that enable devices to discover information about the available roaming partners using query mechanisms. The Access Network Query Protocol (ANQP) ([132]) is the query and response protocol, which supports Hotspot 2.0. ANQP, apart from the operator's domain name, the accessible roaming partners, the type of the access point (private, public free, public chargeable, etc.), is capable of collecting the load information (i.e., total number of currently associated de- vices to the AP, channel utilization percentage and an estimate of the remaining available admission capacity). With load being one of the most crucial parameters linked with the QoS provided by a BS or Wi-Fi AP, this feature is of upmost importance when it comes to the evaluation and selection of the most appropriate RAT.

SaMOG [133] allows UEs to seamlessly handover between cellular and Wi-Fi network. According to SaMOG specification, the Wi-Fi gateway does not connect directly to the EPC via the Packet Gateway (PGW). Another network entity, the Trusted Wireless Access Gateway (TWAG) is used, acting as the perimeter security entity of the EPC network and connects to the PGW over a secure GTP tunnel.

ANDSF [11] is a cellular technology standard -closely coupled with the Policy and Charging Rules Function (PCRF) [134] – that implements dynamic data offload for the User Equipment (UE) in a structured method, while in addition, enables the operator to store its policies for discovery and selection of RATs on a server. The UEs are updated with these polices by the server. The policies within ANDSF contain information on which of the available Wi- Fi hotspots are preferable during a specific time or day, and at a specific location as well, based on indications from past measurements.

The ANDSF Management Object (MO) is the primary representation of ANDSF. ANDSF MO may contain information with regard to the UE location, Inter-System Mobility Policies (ISMPs) and Inter-System Routing Policies (ISRPs) [28]. The ISRPs are available for UEs, which support IP Flow Mobility (IFOM), multiple- access Packet Data Network (PDN) connectivity (MAPCON), or non- seamless offload ([135]–[137]). MAPCON enabled UEs may establish different PDN connections through different RATs. IFOM enabled terminals may establish a single PDN connection via multiple access networks, for instance 3G/LTE and Wireless Local Area Network (WLAN). For such UEs, IFOM enables to move individual IP flows from one access network to another with session continuity. The ANDSF prioritized rules in the case of MAPCON apply per PDN connections, while in IFOM and non-seamless offload cases per flow. ANDSF communicates with the UE over the S14 reference point.

The WLAN_NS working item of 3GPP ([14]) is working to enhance 3GPP solutions for WLAN and access network selection based on Hotspot 2.0 and ensure that data, i.e. Management Objects (MO) and policies provided via HotSpot 2.0 and ANDSF are consistent. This alignment of ANDSF and HotSpot 2.0 provides an excellent basis for the complementarity of ANDSF and Hotspot 2.0, as well a number of multi-operator scenarios that can be supported. In [15], a rather exhaustive list of possible scenarios is presented, where cellular operators and wireless Internet service providers can cooperate and allow UEs to roam among them.

From the analysis above it is clear that current standards and solutions pave the way to the operators to deploy flexible solutions where data flow can be handled, even on a per session basis, using various RATs. Also, much effort has been spent to communicate an operator's RAT selection policies to UEs. Although we have highlighted many key

enablers from the standardization efforts that have evolved recently, there are still open problems that the industry needs to address before Wi-Fi/Cellular integration can be fully realized. These technical enablers on the one hand facilitate the design and development of a new generation of RAT selection mechanisms, – something required for the realization of 5G networks and their high QoS requirements –, however, they need to be federated by novel mechanisms that will bridge the gap between these integration efforts and the intelligent handling of the 5G HetNets. In particular, many of the challenges facing Wi-Fi/Cellular integration have to do with realizing a complete intelligent network selection solution that allows operators to steer traffic in a manner that maximizes user experience and addresses some of the challenges at the boundaries between RATs.

Current shortcomings, such as the static nature of the routing rules that are applied have to be addressed; thus, real-time dynamic RAT selection and traffic steering protocols will further federate the current efforts. ANDSF provides a very useful framework for distributing operator policies, however, there is additional in- formation, which is likely available, and which could be used to improve network selection decisions. Avoiding "unhealthy choices" such as the selection of a Wi-Fi AP with a strong signal but very limited bandwidth or high load, or the choice of a RAT leading to ping-pong effects in cell edges are issues that still need to be tackled.

2.2.2 LTE – Wi-Fi coexistence from the literature perspective

The co-existence of LTE and Wi-Fi has also been studied in the literature; the key inferences identified in such efforts ([138], [139]) also confirm the aforediscussed potentials along with the issues and limitations that accompany the tighter coupling. In [17] the authors describe the numerous degrees, in which the Radio Access Network (RAN)-level integration may improve the LTE/Wi-Fi cooperation in the context of the emerging Heterogeneous Networks (HetNets). At the same time, however, they highlight the fact that novel multi-RAT and multi-tier solutions require additional infrastructure enablers, such as network management interfaces, able to deliver flexible core network connectivity for the envisioned system architecture of next-generation 5G systems. According to the authors in [140] the LTE-Wi-Fi interworking is a promising solution, however, due to its considerable drawbacks (i.e., signal attenuation through walls, CSMA/CA limitations allowing only one link to be active at one time, etc.), supplementary solutions could be considered such as optical wireless communication systems, such as Li-Fi [141]. Still, they also conclude that big standardization efforts are required in order to define new modes of simultaneous transmissions to multiple users, adding also the challenge of the complexity limits with larger number of antennas.

2.3 Context Awareness and Data Analytics towards Network Resource Management and QoS Optimization

2.3.1 Introduction on Context Awareness

Optimized resource management solutions will be of paramount importance in 5G cellular networks. The tremendous increase of constantly connected devices requiring respective resources will inevitably pose significant challenges to the policies, according to which, a network operator handles available resources. This device- and technology-wise heterogeneous environment calls for a novel, holistic and comprehensive framework, which will build on the available contextual information and optimize network resource distribution and RAT placement among users and devices.

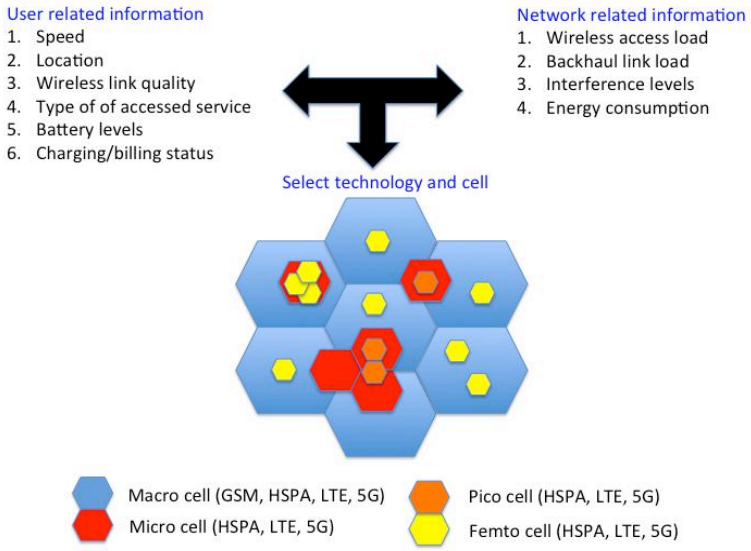


Figure 6: Example Context Information Items for RAT selection in UDNs

User centric information (i.e. all user-related historical context data such as traces, calls e.tc.) can be employed in order to build personalized profiles that adequately capture specific behavioral patterns in terms of mobility, location, services, etc. Data analytics schemes for context-extraction can utilize historical context information related to User Equipment (UE) behavior and produce context-based models and/or UE profiles. Apart from UE-related information, context generated from the core network entities can also be exploited. Network context information processing and sharing is obviously necessary for context-based management and control schemes in the context of a 5G ecosystem.

The various contextual items collected by a UE can be randomly transmitted to a knowledge extraction entity multiple times as required, thus increasing the signaling overhead induced. In parallel, when dealing with historical information coming from UEs, and/or the core network, we expect that a repeated pattern can be identified due to the strong temporal and spatial nature of their activity. This repeated data pattern –when processed by data analytics schemes– typically does not provide additional knowledge to already available models. Evidently, when considering thousands or millions of coexisting devices this comprises a huge overhead in terms of network and computing resources.

In order to support the communication requirements of 5G networks while in parallel facilitate information exchange among network entities for context extraction, novel, sophisticated solutions have to be developed and deployed. These solutions should address data pre-processing prior to transmission by means of aggregations, data reduction, outlier removal or filtering in order to minimize the number of messages as well as the data size. Relevant frameworks should both target on one hand to extract knowledge from network information while –on the other- refrain from excessively increasing the induced signaling overhead. Evidently, building on the derived knowledge, databases of policies and rules that facilitate the mapping of user profile classes to specific RATs and cell layers (macro, pico, femto, etc.) can be populated. Ultimately, the network operators via such tools will be able to maximize the efficiency of the placement of the various UEs, optimizing network resources' provision and distribution, as well as the quality of service and users' quality experience.

2.3.2 Context Data Analytics towards 5G: Current status in 3GPP

3GPP's latest standardization reports move towards proposing the Network Data Analytics module (NWDA) ([154]); NWDA Function (NWDAF) provides slice-specific

network data analytics to the PCF (Policy Control Function). NWDAF provides network data analytics to PCF on a network slice level and the NWDAF is not required to be aware of the current subscribers using the slice. The data may be aggregated from various network elements and functions, such as PCF (Policy Control Function), ANDSF, OFCS, NMS, etc. Based on data analytics methodologies, contextual information with regard to the subscribers' traffic and mobility patterns is extracted and utilized towards resource management optimization.

2.3.3 Context Information Aggregation and Processing: State of the Art

The context-aware optimization of RAT/cell layer selection or other RRM operations in general has been addressed by numerous solutions. Context is typically built using network and/or UE data typically comprising several parameters – derived in principle from historical measurements – such as user location, device capabilities (e.g. CPU, OS, RAM, RAT support etc.), type of consumed service etc.

Besides the definitions provided in the previous section, the context definition is further enhanced in [156] by the concepts of inferred knowledge (apart from the simple understanding of the environment stimuli), context information dissemination among potentially heterogeneous and remote entities, as well as context building based on interactions between users and environment. Apart from the definition of the notion of context, the authors go one-step further and introduce the key functionalities of a Context Aware Mobile and Wireless Networking (CAMoWiN) system; the main functionalities related to a CAMoWiN system are the following (Figure 7):

- *Context Acquisition:* the functionality of acquiring raw data and extracting a first version of context. The raw data are measurements (from the wireless medium). Then cleansing and simple correlations are being performed for extracting the "low level context". The context acquisition approaches should be easy to deploy, easy to use, and non-intrusive for the end users [158].
- *Context Modeling:* the modeling functionality interprets low level context into higher-level context. This is related to derivation of further implications and representation in such way, that may be stored/retrieved/exploited efficiently. The context modeling is responsible for acquiring the low level context and enhancing it (to high level context).
- *Context Exchange:* Context exchange deals with propagation of low and higher-level context information from sources to sinks. In general, it is considered that context that it is not being exchanged is useless.
- *Context Evaluation:* Context evaluation is related to the linking between higher level context to final system actions. The context evaluation shall be dynamic, QoS aware and user oriented. At the same time the context evaluation shall minimize human intervention.
- *Business Logic:* Business logic is the link between final system actions (i.e., Context Evaluation) and the strategic business decisions. Business decisions are in general related to business modeling, service personalization (i.e., consumers' profiling, service impacts, etc.), and even non-technical dimensions (i.e., socio-economical, ethical, psychological etc.).

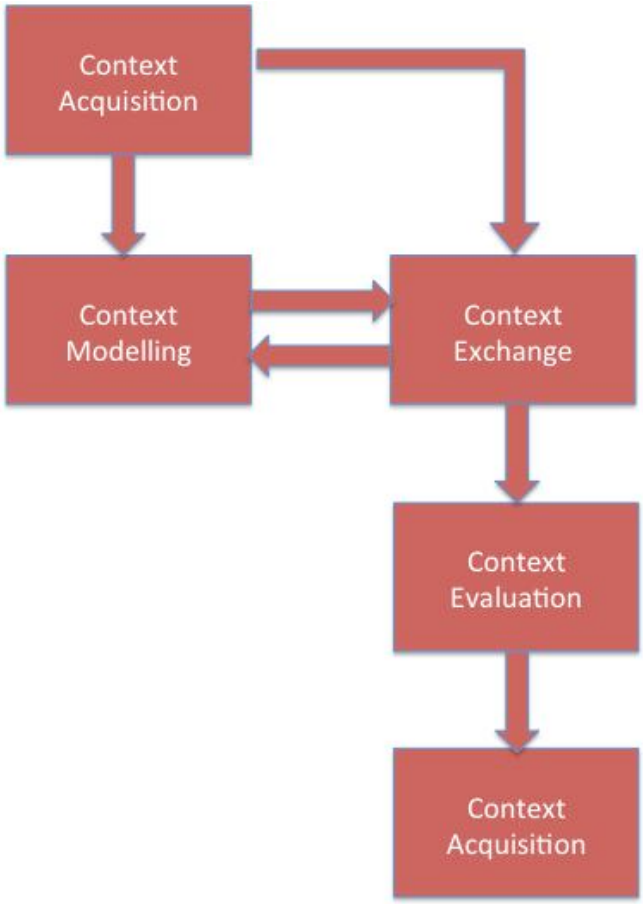


Figure 7: Main functionalities related to Context-based approaches

Similar functionalities and requirements are also considered in [159], [160]. A context-aware framework that assumes all the aforementioned functionalities in order to perform energy aware HO is introduced in [161]. The framework groups the key functions differently however, it assumes the same key operations. The extracted context (measured locally or introduced manually by the use) is based on service inputs (Quality of Service (QoS)/ Quality of Experience (QoE) class, delay tolerance, transmission power rate, battery consumption, etc.) and is not further processed/ enhanced.

In [162] and [163] authors propose a dual-purpose context aware system. The proposed CA system enables sophisticated service provision (UE grouping/multicast) or enhanced RRM (sophisticated handover, admission control, etc.). The first case capitalizes on users' grouping and enables group service provision for efficient resource usage. This approach covers all the CAMoWiN functionalities and exploits the current network view for context extraction. The second case additionally exploits position information for performing proper handover decisions and minimizing the number of unnecessary handovers. It is worth pointing out the proposal in [163] – an idea closely related to our own – where the authors employ data mining techniques in order to cluster users and optimize service provision. However, the proposal does not take into account historic information and confines itself to the current network view. Moreover, the proposed scheme is only capable of grouped service provision (e.g. multicast). In general, however, the context extraction schemes follow simple approaches that are policy-based or objective function based ([161], [162]). A few schemes use more sophisticated approaches for identifying correlations between parameters such as Bayesian networks ([164]-[167]).

From a holistic point of view, the afore-mentioned mechanisms that utilize user behavioral information are primarily based on direct/ online gathering of information and non-optimal

information transmission (i.e., without any pre-aggregation of information, pre-processing, filtering, compression, etc.). As a result, existing approaches result in excessive signaling overhead. Few efforts attempt to combine the context-based knowledge extraction with context preprocessing for signaling minimization between involved network entities. In [168] authors describe a framework that comprises preprocessing primitive context information and transferring it into knowledge via tools like categorization, correction or compression. However, the particular framework remains purely at the data management level, as it does not cope with the problem from the perspective of the network signaling optimization, since the actual information exchanged between network entities is never referenced. Similarly, authors in [169] propose a model-based autonomic context management system that can dynamically configure its context information gathering and pre-processing functionality in order to provide fault tolerant provisioning of context information. The approach aims at increasing openness, interoperability of context-aware systems; however, it does not manage to present an overall solution from the network signaling perspective.

In addition to the academic research efforts, several patents have been claimed related to context information preprocessing mechanisms, compression techniques, redundant traffic reduction, routing optimization methods etc. Due to space limitations, in the context of this paper, we present the most relevant, advanced and indicative solutions. Authors in [170] consider bandwidth limited communication links and evaluate the efficiency of the compression of a communication protocol. Packet header compression is another method ([171]) that minimizes the signaling overhead, however, no context awareness is taken into account, resulting only in minor improvement. Besides compression, information transfer optimization can be achieved by enhancing routing techniques as well. Optimizing mobile traffic data management via optimized polling intervals ([172]) is a paradigm that attempts to enhance the information sharing and minimize signaling overhead in mobile networks. Such methods include batching data that are directed to a mobile device received over multiple transactions, so that a connection is established only once and not for every transaction. However, it does not apply redundant information identification and removal. Another approach to implement redundant traffic reduction in wireless networks ([173]) – especially when a device requests data download from the network – is by identifying another device in the same network that has at least a portion of the requested data. In this way, the initiating device eventually requests the remaining portion, offloading this way the wireless network. Such solutions are based on caches residing on each computing system for making the portion comparison. Finally, as described in [174], the pre-fetching and preparation of certain content may serve as a means for optimizing the information sharing process among network entities. The idea is to store content data replicas in two or more network locations and perform prediction regarding context requests by users and devices. The processing step includes content transcoding in order to ensure compatibility with the predicted user and his device. The method, however, claims no context consistency evaluation or redundant data identification and removal.

To summarize, the afore-mentioned mechanisms attempt to address several aspects of the context information processing, however no holistic, end-to-end solution is currently available. It is clear that existing solutions are either too simple to implement but achieve sub-optimal solutions, or they provide significant improvements but their complexity and the burden – in terms of signaling overhead – placed on the network components makes them unattractive for a real deployment by the operators. It is therefore evident that existing solutions need to cover a significant gap in order to efficiently and realistically support the real needs of 5G networks. Our approach is based on sophisticated data mining schemes and aims at using the context information for extracting user profiles. We

also address the need for minimization of the signaling overhead between network entities, requiring information exchange, via context and user profile-related filtering techniques. The forecasted tremendous increase in the number of devices in the near future, and – as a direct consequence – the resulted excessive signaling cost, make the need of such a solution imperative.

3. RAT SELECTION OVERVIEW: CELL (RE-)SELECTION, ADMISSION CONTROL AND HANDOVER FROM A COMMON POINT OF VIEW

3.1 Introduction

In 5G networks a plethora of 3GPP (GSM, HSPA, LTE, LTE-A) and non-3GGP Radio Access Technologies - RATs - (e.g., Wi-Fi) will be co-existing. The deployment scenarios mainly consist of a multi-layer use of macro, micro and femto-cells. Thus, multi-mode end devices, supporting different applications with diverse QoS requirements, will be served by different technologies. The mapping of end devices to the most appropriate RAT/layer is a complex process. Special care needs to be taken from the mechanisms that will decide and execute this mapping, as - especially in future ultra-dense environments- sub-optimal configuration of the networks elements will affect -on the one hand- specific critical 5G use cases with strict QoS requirements (industrial M2M scenarios with ultra-low delay requirements), as well as the overall performance of the network. Typically, in cellular networks three mechanisms affect for the appropriate placement of end devices to RATs/layers, namely: cell (re)selection, admission control and handover.

Cell (re)selection is a device control operation while admission control and handover (in the case of horizontal handovers) are network controlled operations assisted by the end devices. The cell selection procedure is related to the identification by a UE of the most appropriate cell to associate to, from all the available frequencies of each supported RAT. This procedure must be efficient in terms of the required time to select the most appropriate cell. This requirement becomes even more challenging when the number of cells, frequencies, and RATs to be evaluated increases. Once a UE decides which is the optimal cell for it to be camped in, it starts a process to re-evaluate its situation and possibly to discover a more suitable cell. Such process is named cell reselection. This process starts only after the UE is being camped in a cell for at least a specific amount of time (operator-configurable cell reselection parameter).

Admission control is the procedure of admitting a new incoming session by considering the available resources of the network, in conjunction with the user requirements as well. It should be noted that the admission control procedure is closely related to the transition from the RRC IDLE to RRC CONNECTED mode. When the UE tries to change its state and start transmitting, the net- work will evaluate its current status (as well as the available resources) and decide whether to allow to the UE to change its state and allocate to it the corresponding resources.

Similarly, to the previous mechanisms, ***handover*** is one of the key mobility mechanisms related to network resource management. In the majority of the cases, the decision for handover regarding a specific UE is based on the radio conditions that the UE is experiencing while being in RRC CONNECTED mode (in contrast to the cell reselection that is linked to RRC IDLE mode). According to one of the first 3GPP definitions, the handover process is defined as the process, in which the radio access network changes the radio transmitters or radio access mode or radio system used to provide the bearer services, while maintaining a defined bearer service QoS ([17]).

In order to provide a comprehensive overview of the UE placement process to the respective RATs and cell layers, we consider as extremely useful to study all three afore-mentioned mechanisms following the similar methodology and from a common point of view. The challenges, which stem from the forthcoming 5G use case requirements influence to a great extent all three procedures, as the management of the network resources in the future heterogeneous networks of 5G will have a critical impact both on the Quality of Service and Experience for the users, as well as the efficiency, safety and

scalability of critical uMTC or mMTC use cases. Furthermore, the outcome of each one of these mechanisms' decisions on the network operation has a direct impact on the other two, as all are directly linked to the selection of the optimal RAT or cell layer for a specific UE with specific QoS requirements. As a result, the need for a framework, which ultimately acts on top of all three RAT selection mechanisms, having the decisions' and policies' overview is imperative.

As we will present later in this section, existing standards rely mainly on a set of dynamic parameters like the signal strength, the network load, the current mobility of a user, etc. More sophisticated research proposals take into consideration inputs like the user preferences, although it is not clear how these preferences are defined. Other proposals suggest using real time and past information such as the location and speed of a terminal, the type of service, the experienced QoS, the available bandwidth, the energy consumption, the user profile, etc. All this information is referred as context information and it has been shown to improve the network performance and eventually the quality of experience for the users while satisfying the network operator policies. Context information will be analyzed in the next section of this thesis. All these mechanisms use appropriate tools (e.g., utility functions, fuzzy logic, etc.) to evaluate the context information as a multi-criteria decision problem and reach a decision.

Diverse information is used by the three primary RRM mechanisms. In general, they use some type of contextual information as input to a clearly defined strategy. For example, upon identification of a fast moving UE, it should always be placed in macro cells in order to avoid unnecessary handovers. Or, consider the case of users that are mainly/ only accessing voice services; these subscribers should be placed in legacy RATs (e.g., 2G) so as to reduce the signaling and data load of an LTE access network for more demanding users. On the other hand, these mechanisms should meet the operator's policies for reduced signaling, low blocking probabilities, energy savings and increased capacity.

3.2 Cell (re-)selection

3.2.1 Cell (re-)selection procedures according to 3GPP

In the 3GPP specifications ([18]), the cell selection procedure is based on the S-criterion which is linked to two parameters, namely the Reference Signal Received Power (RSRP) and the Reference Signal Received Quality (RSRQ) that indicate the link quality level. For avoiding ping - pong effects, the sensed values of the target cell shall be higher of the corresponding values of the cell where the User Equipment (UE) is camped plus an offset. In order to make this procedure more efficient, the UE searches only the RATs of a specific Mobile Network Operator (MNO) and it may also exploit information from previous accessed cells. It should be added that when a UE turns from connected to idle mode, it may not remain camped at the cell where it was connected. This could happen because the connection release message can include network information directing the UE to search for a cell on a particular frequency.

In conjunction with the cell selection, the cell reselection takes place by the UE in order to re-evaluate its situation and possibly to discover a more suitable cell. The cell reselection procedure is based on a set of requirements and evaluations. Such evaluations are related to the following aspects:

- Measurement Rules: measurement rules are applied for enabling the UE to save battery power. The rules define that the UE shall perform intra and inter frequency measurements when the quality of the serving cell is below or equal to a predefined threshold, minimizing this way the energy that is consumed for the measurement process.

- Frequency/ RAT evaluation: the network may propose to the UE frequency/ RAT priorities for performing cell selection and reselection. This implies that the network sends to the UE proposals on the frequency that is more preferable to associate. The UE, when the measurements of the higher pri- ority frequency is below a certain threshold for a time-window then the UE starts evaluating other frequencies as well.
- Cell ranking: the cell ranking of the same frequency (intra-frequency cell selection) is being performed using the R-criterion. Similarly, to the S- criterion, the measured RSRP or RSRQ value is being augmented using an offset -provided by the network- so as to avoid Ping-Pong effects. Addition- ally, considering the user mobility, the offset may be increased by specific values, so as to make some cells unattractive for highly moving users.
- Accessibility aspects: Accessibility aspects are related to cell access restrictions by a closed subscriber group (CSG) in a small cell (i.e., macro, femto, pico cell) or even access to a cell due to emergency or priority services.

3.2.2 Related papers

In the literature, several approaches have been proposed regarding cell selection and reselection for future cellular networks. These solutions attempt to improve the performance towards two directions, namely: a) on avoiding useless handovers, when the user initiates a service in an unsuitable cell and b) on minimizing the probability for a UE to receive a call admission control reject when it initiates a new session.

In general, the cell selection and the cell reselection are based on reference signal received power (RSRP) or reference signal received quality (RSRQ) measurements. It should be noted that RSRP and RSRQ bias is something typical in such types of algorithms for avoiding ping - pong effects. The cell (re)selection procedure may be either centralized (taking place primarily on the network side) or decentralized (mainly on the UE side). In the majority of the surveyed schemes ([19]-[33]), the authors describe decentralized solutions. Several of the aforementioned mechanisms ([19], [22], [23], [26], [30], [33]) utilize contextual information to facilitate the cell selection decision. Such contextual information comprises historical information, current capacity of the backhaul network, bandwidth of the base station, etc. In some of the mechanisms, projections and probabilities are also being used ([19], [26], [30], [33]), for instance, in order to estimate expected data rates for particular QoS class traffic flows. Furthermore, some of the mechanisms focus on interference management, either by interference mitigation techniques ([21], [27], [28], [31]), Almost Blank Subframes (ABS) techniques ([22], [29], [32]) or power control schemes in general ([21], [25], [26], [31]). From a more general perspective, all proposed schemes are taking into account some metric related to the received signal strength. Indicatively, numerous proposals ([19], [26], [27], [28], [30]) utilize the SINR metric. Other schemes ([20], [21], [25], [29]) use the RSRP metric without bias, while several others are using bias factors as well ([20], [29], [30], [30]). In parallel, several mechanisms integrate more complex methods in their solutions or validation, typically combined with the afore-mentioned techniques and metrics, such as objective functions ([22]), game theory ([26]) and ergodic theory ([27]). The afore-presented literature proposals are further elaborated in the following paragraphs.

Some mechanisms have proposed the Range Expansion (RE), a typical RSRP with bias scheme ([19]-[26]). After experimenting with several bias levels it is reported in [19] that even without the bias, a significant number of users associate to the small cells, whereas, as the bias level increases, the cell spectral efficiency decreases. Similarly, in [20], in an extension of the previous scheme; instead of associating a UE with the cell with the

highest RSRP augmented with the bias value, the authors propose to form a priority list consisting of all the cells that are above the predefined RSRP thresholds and camp to the first small cell of the list. This approach should be complemented with an Inter Cell Interference Coordination (ICIC), in the presence of which, the cell edge spectral efficiency is being significantly benefited. The introduction of the priority list further benefits the spectral efficiency of the system, because the users that are close to the small cells prefer to associate with them, but only if their basic service requirements are being satisfied (above a certain threshold). Another Range Expansion approach is proposed in [21] where the authors propose a small threshold of 1dB. Their main innovation however lies in the incorporation of different types of small cells; incorporated small cells are pico cells (i.e., operator deployed) and femto cells (i.e., user deployed). Their conclusion is that users associated to small cells with unlimited backhaul capacity are significantly benefited. Three schemes of cell selection, namely, the max RSRP scheme, the RE scheme and the expected data rate scheme using ABS are being considered and compared in [22]. ABS are sub-frames without any activity or only transmitting the reference signals without any types of control or data signals ([23]). In the RE scheme, they experiment with both static and dynamic offset setting. The simulation results show that when the offset value of the cell range expansion increases, more UEs are offloaded to pico cells even if they are far from the pico cells. Range expansion is also studied in [24] combined with techniques such cell splitting, range expansion, semi-static resource negotiation on third-party backhaul connections, and fast dynamic interference management for QoS via over-the-air signaling. RE scheme combined with bit rate probability estimation as a function of the ABS is studied in [25]. Simulation results show that the performance of the proposed cell selection scheme improves the average user data rate and increase the number of offloaded users as the ABS ratio increases. In [26], the exploitation of information related to shadowing that impacts cell selection is presented. A cell selection scheme for cell selection exploiting downlink data rate probability coverage and Range Expansion is once more shown. In addition, long term shadowing is being proposed in order to avoid using instant calculations. Almost Blank Sub-frames are being considered also in [27], however, in this paper, instead of the Range Expansion method, the cell selection procedure is based on an objective function, which takes into account as parameters the available ABS and the available bandwidth. The simulation results show that the multiple objective function for cell selection leads to significant gains in the average downlink capacity and the number of associated users.

The received signal strength and quality-related metrics, i.e. SINR, RSRP and RSRQ are widely used in the majority of the proposed schemes ([26]-[32]). In [28], the authors experiment with all three metrics associated with ICIC in heterogeneous networks in the LTE-Advanced (LTE-A) downlink. Simulation results show that the downlink cell and cell-edge user throughput levels of the RSRP based cell selection are degraded compared to those of SINR-based cell selection. A (decentralized) SINR based method combined with interference avoidance schemes for MIMO mode selection in multi-cell environments is proposed in [29]. The scheme is validated using ergodic theory and the simulation analysis shows that the proposed scheme outperforms typical schemes based on Fractional Frequency Reuse (FFR) with regards to spectral efficiency. An enhanced cell selection scheme that considers scheduling opportunities available at the femto cell is analyzed in [30]. The main purpose of this solution is to utilize the underutilized resources of the femto cell. The scheme in [31], is based on the concept of Dynamic Cell Selection (DCS). According to this, when a user is in a multi-coverage region, he may have opportunities to select another link better than the serving sector. In the DCS enhanced with ABS so as to mute the interferers, muting is applied to the cell with the second highest instantaneous received power for decreasing the interference to cell-edge UEs. In [32],

the problem of cell selection and resource allocation in heterogeneous wireless networks is being investigated and formulated as a two-tier game. Specifically, the authors propose a distributed cell selection and resource allocation mechanism, in which UEs perform the cell selection and resource allocation processes independently. Finally, in [33], the authors propose a cell selection and reselection scheme for several tier networks; each tier has different path loss model. According to the proposed scheme, a pico cell may claim a user even though he has stronger macro cell signal because the user "connects" to the (H)eNB that offers the maximum long term averaged received power (with biasing).

The following table summarizes the findings of the performed state of the art analysis. More specifically, we observe that all the methods consider RSS measurements in several ways, namely pure RSS, and RSRP. In general, bias in favor of the small cells is used. In some cases, the expected bit rate is calculated using the Shannon capacity formula. Furthermore, all but one approaches, are decentralized/UE-oriented. The one approach that it is centralized exploits the UE RSRP measurements and proceeds in power control for interference management. Most of the methods assume collaboration with interference management/coordination schemes. The interference management/coordination methods in some cases are part of the cell selection and reselection. ABS and power control are the most popular interference management schemes. The ABS schemes are also used for the identification of the available/expected throughput. Regarding the use of the available context information, in general the exploitation is poor. In some cases projections are being done with regards to available/expected throughput in a cell. In two cases the backhaul link and the available BW of a cell are considered.

Besides the plethora of solutions that was presented above, in the literature, no case has been identified to our knowledge that exploits detailed past information about a user habits (e.g., specific cells/RATs in a location/time/day) and needs (which kind of service usually the user requests). The exploitation of such information could be proven beneficial especially in cases where an operator has deployed a multi-RAT network, since users tend to repeat similar behaviors with regard to mobility patterns, service and application usage, etc.

Table 1: Cell (re) selection mechanisms state of the art overview

	Centralized/ Decentralized		Available Context				Interference Management			RSS			Throughput-QoS		Method – If Complex
	Centralized	Decentralized	History	Projection/ Probability	Backhaul	Available BW	Interference Mitigation	ABS	Power Control	RSRP	RSRP+ BIAS	SINR	Expected Bit-rate	RSRQ	
[19]		X		X								X	X		
[20]		X								X	X				
[21]		X					X		X	X					
[22]		X				X		X					X		Objective function
[23]		X			X								X		
[25]	X								X	X					
[26]		X		X					X			X	X		Game

[27]		X					X					X			Statistical & Ergodic
[28]		X					X					X		X	
[29]		X						X		X	X		X		
[30]		X		X							X	X			
[31]		X					X		X		X				Objective
[32]		X						X					X		
[33]		X		X									X		

3.2.3 Related patents

The benefits from proper cell selection and re-selection has attracted the interest from researchers and industrial organizations to claim patents. Thus, several patents have been proposed and introduced in this scientific area. Similarly with the efforts presented from the literature, mainly the focus is in simple RSS/RSRP/SINR based solutions using bias ([34], [35], [37]-[39]). The decisions in all cases are performed in the UEs, based on local view as well. A UE makes its decision based on its own measurements. Furthermore, contrary to the research papers, where the proposals are in general related to interference management/mitigation schemes, the patents do not correlate cell selection/reselection to (e)ICIC. Moreover, it is worth mentioning that several mechanisms are focusing only in cell reselection, assuming the presence of a cell selection scheme (RSS/SINR based). Regarding context, the proposed patents only focus on the available bandwidth in the target cell, or the probability of being served by a specific cell. The following paragraph provides some further details in relation to the analyzed patents:

In [34], a method for performing a selection or a reselection of a cell in a UMTS or LTE network, based on a parameter taking into account the traffic load level of the cell is claimed. The cell selection and reselection procedure is a decentralized UE decision and is based on (link) quality measurement for all cells and the load information (such information is being broadcasted). In [35], the authors propose a cell re-selection scheme using a ranking algorithm. The method is located at the UE side and is based on priority-based selection of cells of the same or different technologies. The criteria (i.e., RSRP, and received signal code power (RSCP)) used for the selection depend on technologies that are available in each case. The scheme also proposes an adaptable offset so as to minimize the ping-pong effects. In [36] it is proposed to apply prioritization on the selection of the HeNBs over eNBs. The following approaches are proposed: a) the use of SIBs for the transmission of differentiation information, b) the use of different Location Area Codes (LAC) for eNBs and HeNBs, c) the use of previous knowledge regarding the layer of the eNBs, d) the use of specific bit sequences (scrambling codes) for differentiating the layers of the eNBs. In [37] the authors claim an SINR based scheme, using bias for small cells. According to this approach, all carrier frequencies may be directly compared, using different bias margins, depending on the technology (i.e., different adaptability thresholds for each technology). An SINR cell (re)selection scheme is proposed in [38]. This scheme takes into account Mobility Management aspects, -and more specifically-, in cases where the UE sensed RSRP from the candidate cell is not significantly higher than that the RSRP of the camped cell, the UE sends Tracking Area Update (TAU) messages and in case the Mobility Management Entity (MME) replies in the micro cell then the UE is static so it should be camped to the microcell. In [39] the authors claim a cell reselection method in heterogeneous networks. The method relies on an adaptable threshold, based on the resources used for Cell Reference Signal (CRS - cell-specific reference signals transmission). If the same resources are used, then a negative bias is used to make the candidate cell less preferable, whereas if different resources are being used, a positive bias makes the candidate cell more preferable. Finally, in a network and cell selection and reselection for LTE inter-radio access network based on mobility drivers is claimed.

Ten mobility drivers have been defined in the 3GPP Technical Specification TS 36.300 ([40]), namely, best radio condition, camp load balancing, traffic load balancing, wireless device capability, hierarchical cell structures, network sharing, private networks/home cells, subscription-based mobility control, service-based mobility control, and Multimedia Broadcast Multicast Services (MBMS). These mobility drivers are targeted to be optimized by a centralized decision entity. Additionally, typical inputs for cell selection and reselection, such as the RSRP and RSRQ are being used. The centralized decision entity, exploits the overall network view that it is available, as well as the knowledge of the device and eNBs capabilities and decides the optimal camping (and placement) of the devices in the network.

In Table 2, which follows, the reader may easily identify where the researchers have turned their interest up to now. It must be highlighted that despite the diverse methodologies that have been applied, -both in literature, as well as in the listed patents-, what still has not been identified is a solution that relies on a greater extent on past context information or profiling of the UEs. The 5G environments will introduce much more complex ecosystems with multiple co-existing RATs, cell layers, UE types and traffic patterns. A solution that builds upon this past context information, extracts UE profiles and uses them in order to optimize –even on a personalized level- the policies of the network with regard to RAT selection and UE placement is of high value.

Table 2: Cell (re-)selection patents overview

	Centralized/ Decentralized		Available Context				Interference Management			RSS			Throughput-QoS		Method – If Complex
	Centralized	Decentralized	History	Projection/ Probability	Backhaul	Available BW	Interference Mitigation	ABS	Power Control	RSRP	RSRP+ BIAS	SINR	Expected Bit-rate	RSRQ	
[34]		X				X								X	
[35]		X									X				RSCP
[36]		X	X												SIBs, LAC, scrambling codes
[37]		X										X			SINR+ bias
[38]		X		X								X			TAU
[39]		X									X				

3.3 Admission Control

3.3.1 Admission Control procedures according to 3GPP

As already discussed in the previous section, call admission control is the procedure of admitting a new incoming call/session by considering the available resources of the network, in conjunction with the user requirements. Call admission control procedure, -being closely linked to RRC IDLE to RRC CONNECTED UE modes-, involves the RRC connection establishment procedure that can be triggered by either the UE or the network. For example, the UE triggers RRC connection establishment if the UE moves into a new Tracking Area and has to complete the Tracking Area Update signaling procedure. The network triggers the RRC connection establishment procedure by sending a Paging

message. The flat network architecture of LTE removes the requirement for these signaling procedures. In the case of LTE, the initial Non-Access Stratum (NAS) message is transferred as part of the RRC connection establishment procedure. In the case of UMTS, the initial NAS message is transferred after the RRC connection establishment procedure. The approach used by LTE helps to reduce connection establishment delay. RRC connection establishment configures signaling Radio Bearer (SRB) 1 and allows subsequent signaling to use the Dedicated Control Channel (DCCH) rather than the Common Control Channel (CCCH) used by SRB 0. The entire procedure is completed using only RRC signaling. The RRC Connection Request message is sent as part of the Random Access procedure. There is no scope for the UE to report any measurements within the RRC Connection Request message. The UMTS version of the RRC Connection Request message allows the UE to report CPICH measurements, which can subsequently be used for downlink open loop power control calculations. In the RRC connection request the establishment cause is included; the cause may be one of the following ones:

- Mobile originated signaling,
- Mobile originated data,
- Mobile terminated access (paging response),
- Emergency.

The UE starts the “T300 timer” after transmitting the RRC Connection Request message. LTE uses the T300 timer to define how long the UE waits for a response to the RRC Connection Request message. The establishment procedure fails if T300 expires before receiving an RRC Connection Setup message. The procedure also fails if the UE completes a cell re-selection prior to receiving the RRC Connection Setup message. This increases the delay associated with connection establishment but does not cause the overall procedure to fail unless the maximum number of preamble transmissions has been reached.

The decision whether to accept or not an RRC connection request lies on the eNodeB; whether it has available resources or not and is implementation-specific. In case the eNodeB rejects an RRC connection request, it sends an RRC Connection Reject message, which includes a wait time. This is in contrast to the equivalent UMTS message, which also includes a rejection cause, although the UMTS rejection cause can only be defined as congestion or unspecified. The UMTS message can also include redirection information to direct the UE towards another RF carrier, or Radio Access Technology (RAT).

Upon receiving an RRC Connection Reject message, the UE starts the T302 timer with its value set equal to the wait time. Access Class barring for mobile originating calls, mobile originating signaling and mobile terminating access is applied until T302 expires, i.e. the UE is not allowed to send another RRC Connection Request for those connection types, and to the same cell, until T302 expires. T302 is stopped if the UE completes cell reselection. In that case, the UE is permitted to send an RRC Connection Request to the new cell.

3.3.2 Related papers

The topic of call admission control is one of the best studied in the literature of networking. In this section, we provide the key findings and view of industry through existing patents. The survey in [41] proposes a classification of user-based call admission policies in cellular networks. According to the authors, a good CAC algorithm must have the following features in order of importance:

1. Maximize the channel utilization in a fair manner to all calls,
2. Minimize the dropping probability of connected calls,
3. Minimize the reduction of the QoS of the connected calls and
4. Minimize the blocking probability of new calls.

It is proposed that CAC polices may be divided into different categories depending on the comparison basis, namely user (number)-based CAC and interference-based CAC policies. In [42] a new categorization is presented: prioritized, non-prioritized and optimal policies. In non-prioritized policies, all calls are accepted when the requested channels are free; in prioritized, one group of calls may have higher priority than other groups and some calls may be queued or rejected when the requested channels are not available. Optimized CAC policies accept or reject calls in order to maximize the throughput of the network. In [43] the authors propose a different categorization method based on deterministic/stochastic guarantee aspects, distributed/local control and adaptivity to traffic conditions. In this survey, several CAC schemes are compared in terms of performance and complexity; the common characteristic of all these schemes is the handover prioritization; i.e., providing priority to on-going sessions that request a handover, rather than new sessions that are being initiated. The authors consider both optimal and non-optimal schemes and evaluate schemes designed for multi-service networks, hierarchical systems, complete knowledge schemes, as well as schemes using pricing for CAC.

Further, according to [44], in the CAC schemes there is significant need for:

- Use of realistic (non-exponential) mobility and traffic models (packet-based). Novel mobility models should not preserve necessarily the Markovian property. On the contrary, a more accurate description of traffic dynamics is targeted via new traffic modeling techniques. Self-similarity ([45]) is one of the recent findings in traffic analysis that should be taken into consideration.
- Application of cross-layer design in order to improve the performance of CAC schemes and achieve bit-level, packet-level and call-level QoS.
- Design of CAC schemes for multiple services networks in order to support multimedia services, by efficiently sharing the wireless medium.
- Consideration of heterogeneous networks in order to achieve global roaming and end-to-end QoS. This implies that a CAC scheme must be able to communicate with other control components of the network through standard mechanisms to provide end-to-end QoS guarantees.

We analyze a number of CAC mechanisms so as to highlight the outcomes of the aforementioned observations and identify the gaps in the state of the art. From an overview perspective, the majority of the proposed schemes apply the admission control procedure on the level of the eNodeB ([45], [46], [48], [49], [50]). The rest ([47], [51]) have a more generic overview of the network environment when admitting or rejecting new UE sessions. In addition, most of the schemes utilize some kind of contextual information. In particular, the bandwidth seems to be the most popular option, as the majority of the schemes ([45]-[48], [51], [52]) take into account the available bandwidth when making a decision. The latter two combine the bandwidth parameter with additional contextual information, such as the type of the service request ([51]), or the policies of the network ([52]). An interesting point is that two of the proposed solutions link their implementation to the handover mechanism ([52], [53]). More details regarding the afore-mentioned mechanisms follows.

Markov chain-based solutions are usually found in literature ([45]-[47]). A higher-order Markov chain based performance model for call admission control in a heterogeneous wireless network environment is proposed in [45], while in [46] and [47] the authors propose Joint Call Admission Control (JCAC) algorithms, modeling their described policies based on multi-dimensional Markov chains. The afore-mentioned solutions target to reduce the call blocking probability by applying fairness models among mobile terminals (single-mode, dual-mode, triple-mode etc.). Also in [48], the authors propose a CAC scheme for services with unfixed-size data packet or non-periodic transmission (known as non Unsolicited Grant Services – non-UGS) focusing on providing fairness among services, with UGS arrival kept constant. Fairness is achieved through reservation of bandwidth when VoIP traffic exceeds a certain threshold. A CAC scheme with different admit policies for the CSG (Closed Subscriber Group) members is proposed in [49]. The proposed model is based on resources reservation of a part of the available BW only for the CSG members. The rest of the resources are allocated for both CSG and non-CSG members. Fuzzy-logic (FL) is a methodology that has also been applied in past efforts ([50]). FL is used in order to generate the optimal quantity of the channel thresholds so as to assign radio resource efficiently and guarantee the Quality of Service. The investigated FL approach is investigated based on real-time measurements and is a threshold-based CAC combined with the FL policy. Furthermore, a QoS-based CAC mechanism is presented in [51] for avoiding resource overloading in femto cells, combining traffic policing and traffic shaping; the solution is based on the quality of the on-going voice calls passing through the HeNB GW.

As already mentioned earlier, there have been also proposals about correlating the call admission control policies with the handover ones. In [52], the authors emphasize the correlation that should be built between vertical handover decisions and call admission control policies. Using an analytical model, they evaluate the impact on vertical handover by the dropping probability of handover calls derived from the CAC policy in heterogeneous wireless networks. The proposed algorithm, -which integrates CAC policies as well-, is claimed to maximize network utilization while meeting user satisfaction. Similarly, in [53], the algorithm that is proposed is a mobility-aware solution (MA-CAC) with a handover queue (HQ) in mobile hotspots. As it is described, different CAC policies are applied, depending on the vehicle mobility. An analytical model is described in this research approach. To verify it, the authors present simulation results from runs on an event-driven simulator. It is demonstrated, that the dropping probability decreases, while maintaining high channel utilization.

The previously described solutions are being summarized in Table 3. In general, the approaches focus on base stations (BSs) located solutions without specifying the layer (macro/micro/pico/femto cell) or the RAT that it is being used. A few schemes are exploiting the greater network view for making admission control. Furthermore, we observe that all the schemes are considering the available bandwidth, either for making admission control based on the currently available bandwidth or (using projections/estimations) the bandwidth availability in the future. Also a few schemes are considering special admission control strategies based on service consideration inputs; the user context may also be exploited for CAC (mobility measurements). Finally, we observe that a set of mechanisms link the CAC with the HO, which also highlights that these two procedures are closely related to each other; in other words, efficient CAC will minimize the number of the HO.

What it should be mentioned and it is apparent from the summary table is that the previous knowledge is not being used for proper admission control. This is of course a very rational choice since the CAC problem simply focuses on how services requirements can be serviced by the available network resources. However, we argue that the predicted user

behavior, based on the analysis of past information could be beneficial for efficient CAC since a requested service can be accommodated in the appropriate RAT or layer so as to minimize any subsequent HOs or even reserve resources in a RAT or layer so as to support an anticipated future traffic. In any case, Table 3 illustrates that this information is not being extracted or exploited in any way.

Table 3: CAC schemes state of the art overview

	View		Context					Resource reservation	Layer	RAT	Fairness	Link to HO
	eNB	Generic	History	Mobility	Service	BW	Policies					
[45]	X					X						
[46]	X					X					X	
[47]		X				X						
[48]	X					X		X			X	
[49]	X							X				
[50]	X				X			X				
[51]	X	X			X	X						
[52]						X	X					X
[53]				X								X

3.3.3 Related patents

The Call Admission Control and its apparent benefits to users' QoS have attracted the interest of researchers and industrial organizations for claiming patents, apart from producing literature proposals. A significant number of patents is analysed below. We have tried to cluster the different mechanisms according to their architectural implementation characteristics (centralized, decentralized), the contextual information that they take into account for the decision-making, as well as other input parameters, such as QoS classes, bandwidth, etc. Link quality-related parameters are also considered, such as CQI, SINR, PER, etc.

Most of the claimed patents are based on centralized mechanisms ([54]-[56], [58], [59], [61]-[67], [69], [72]-[80]). In addition, several types of context information is evaluated, such as historical data ([67], [68]), location ([60], [77]), while one of the most popular types of context is the service type, which is used by [57], [59], [71], [72], [74], [78], [80]. Almost all CAC schemes evaluate the available bandwidth of the respective base station, while fewer consider in addition the bandwidth of the backhaul as well ([54], [55]). Link quality-related metrics for the CAC mechanisms are found in almost claimed patents as well. Examples of such metrics are general QoS provision ([54][55]), QoS classes ([55], [56], [77]-[79]), CQI metric ([62]-[64]), SINR ([65], [66], [77]) and PER ([58], [72], [73]). In the following paragraphs, additional insights are provided for some of the most indicative cases.

In [54], a CAC scheme that evaluates the available backhaul (wired) link and decides whether to admit the call or not is described. The authors propose a CAC scheme where a user requests to initiate a call via femtocell. The focus of such scheme is to support a minimum Quality of Service for the incoming call. In [56], the authors claim a system for providing call admission control for Long Term Evolution (LTE) network resources shared between a number of user classes. The resources are provided to the users –if they are available- exploiting the policies and the aforementioned user classes. In [59] VoIP

networks are considered; a CAC trying to achieve the best link utilization is proposed. Various points in the network are polled to generate a link utilization parameter. Parameters denoting threshold values of link utilization in the network are compared to the link utilization parameter, while in the context of the proposed algorithm, comparisons are made between the current link utilization parameter and the candidate links' respective parameters. In [60] a CAC scheme based on a central CAC point is proposed. The scheme focuses on BW water filling, by taking into consideration location, maximum bandwidth allowed, allocated bandwidth, per call bandwidth, currently used bandwidth, and a low bandwidth threshold. In [61] they consider the CAC problem as a scheduling problem for both the downlink and the uplink. Compared to the rest of the BW based schemes, in [62] a logarithmic function BW for resource reservation depending on the users' number is being claimed. In [64] the authors propose a call admission control scheme, which exploits the radio channel quality (Channel Quality Indication – CQI) information from several BS. Then the UEs are being admitted based on the target transmission data rate, the contract type and the terminal type.

A set of schemes relate admission control with interference management or interference metrics/KPIs. One of such interference metrics is Interference Signal Code Power (ISCP). In [65] ISCP is used, taking into account in addition the load of the target cell and the neighbouring cells, while in [66] ISCP is combined with the carrier power of the downlink.

Respectively, a group of techniques link CA with power control and management ([68]-[70]). More specifically in [67] an apparatus for call admission control based on transmission power of base station is claimed. The decisions are based on past samples of total base transmission power. In [68] a patent for estimating the load of a call and the impact that this call will have to the network is being proposed. The patent is based on both measured and historical performance data. In [69], the authors focus on transmission power adjustments for the shared channel, in order to satisfy a predetermined target transmission date rate. In [70], they are exploiting load thresholds for performing CAC; the BSs are responsible for identifying overload events and setting the ratio of transmission power over maximum transmission power.

A set of mechanisms is moving towards service admission control instead of CAC ([71],[72]). This is realized per service using typical water filling considering the available BW; in some cases, focusing on specific type of cells, such as femto base stations ([72]).

Some patents are assuming projections for predicting the network performance and considering the currently available resources, by predicting for example the effect to KPIs, such as the frame error rate (FER) ([73]) the delivered QoS ([74]), the outage probability ([75]), or potential congestions by estimating the future users' transmission bandwidth ([76]).

Several mechanisms are trying to perform CAC by identifying the significance of the incoming call. Classification of the users that make the call is proposed or characterization/priotization of the calls; such schemes are linked to policy-based approaches for handling the classes. The decision may be based on several parameters, such as the time up to completing the communication, the transmission data amount, the power required by communication equipment for performing the communication, the interference amount caused on other on-going calls, the location/speed, etc. ([77]). Furthermore, prioritization of call classes may be created based on previously induced policies ([80]).

Table 4 summarizes the findings of the survey in available patents up to now. Initially we observe that in most of the cases the CAC is performed by the BSs exploiting their local view (BW, backhaul link, location etc.) and local projections. We observe that several schemes assume user classes (QoS classes) or QoS requirements of the users. In the

latter case, the QoS requirements of the users are extracted by the service that the user tries to admit, or by rough BW requirements also posed by the users (this also may be extracted by service the user tries to admit). In the former case, of the QoS classes, the CAC mechanisms categorize the UEs to high/low priority or their service admission requests to high/low priority in a static manner (i.e., previously defined classes). In general, all the schemes take into account the local load, which is in general expressed as consumed BW, number of associated users, etc. Additionally, the CAC schemes try to evaluate the quality of the link between the UE and the CAC control point by evaluating the link quality (i.e., CQI, SINR, Received Power etc.). Furthermore, it should be highlighted that all the CAC schemes (with three exceptions) do not refer to specific BS layer (i.e., macro, pico, femto). One single patent links the CAC to its effect to HOs that may take place in the future. Finally, two mechanisms are exploiting previous knowledge, which however is used only for making BW requirements projections (i.e., how much BW a service call may require in the future) and do not exploit the generic user behavior.

Table 4: CAC patents overview

	Centralized/ Decentralized		Available Context					BW			#users	Policies	QoS general provision	Link Quality				Layer	RAT
	eNB	Generic	History	Probability	Mobility	Location	Service	BW	BW Backhaul	Resource Reservation				CQI	SINR	Power Control	PER		
[54]	X								X				X					femto	
[55]	X								X				X						
[56]	X							X				X	X						LTE
[57]		X					X	X				X							
[58]	X							X			X	X					X		
[59]	X						X	X											
[60]		X				X		X											
[61]	X							X											
[62]	X									X	X			X					
[63]	X			X				X						X					
[64]	X			X										X					
[65]	X														X				
[66]	X														X				
[67]	X		X								X							macro	
[68]		X	X	X				X											
[69]	X	X		X				X								X			
[70]				X				X								X			
[71]							X	X											
[72]	X						X						X				X	femto	LTE
[73]	X										X						X		
[74]	X						X						X						WLAN
[75]	X			X				X											
[76]	X			X				X											
[77]	X	X				X		X					X		X				
[78]	X						X						X						

[79]	X							X			X			X				GSM
[80]	X						X	X				X	X					

3.4 Handover

3.4.1 Handover procedures according to 3GPP

Fast and seamless handover is one of the main goals of all cellular systems, but for LTE this is even more significant, as it is characterized by a distributed and "flat" architecture, which consists of only one type of node, the eNodeB. The Evolved Packet System is purely IP-based, meaning that both real-time, as well as data services will be carried by the IP protocol. The LTE access network is just a network of base stations (eNodeBs, eNBs), leading, thus, to this afore-mentioned flat architecture. No centralized intelligent controller is found, while the eNBs are interconnected via the X2 interface with each other, while towards the core network, the S1-interface is utilized. By distributing the "intelligence" among the LTE eNBs the connection set-up time is shortened, while the handover delay is also decreased. It is obvious, that the User Experience is influenced up to a large extent from the efficiency of the handover mechanism. The handovers in LTE are characterized as "hard" handovers, meaning that there is short service interruption during the handover procedure. However, 3GPP supports different types of handover depending on the systems under consideration or even the different alternatives supported for LTE/SAE.

A handover may be from an LTE eNB to another eNB, to a femtocell (HeNB), or even to a completely different RAT, within 3GPP standards (e.g. UTRAN technology network), or out of the 3GPP specified technologies (e.g., Wi-Fi). The second type of handover is characterized as vertical handover.

Another categorization is related to trusted and non-3GGP trusted access. The word "trusted" mostly refers to the trust by the operator. In the case of a Wi-Fi network for example, a trusted network would be one, which is owned and/or managed by the operator of the cellular system. For Intra RAT (LTE) handover, they key network entities are the Mobility Management Entity (MME) and the Serving Gateway (S-GW) ([81]) (Figure 8).

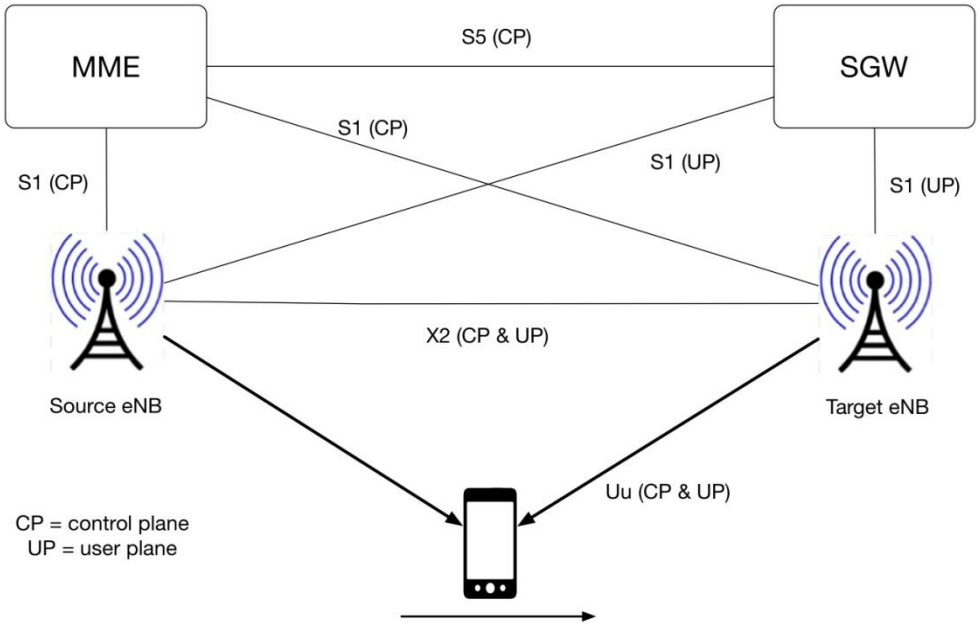


Figure 8: Network interfaces for intra-LTE handover

In case of extreme conditions (e.g., MME overload), the users can be relocated after there

is a trigger from the overloaded MME. Multiple entities inside the network can be relocated and update each other with new information, with this very much depending on the type of the handover that is performed. LTE handover is performed in two main phases: preparation and execution phase.

In LTE, as mentioned before, there are two main modes of operation for a UE at a specific time: “RRC_IDLE” (i.e. idle) and “RRC_CONENCTED (i.e. active) mode. While in active mode, the UE has an active signaling connection and one or more active bearers and data transmission may be on-going. When there is a cell that is considered to be better than the serving cell, and in order to limit interference and provide the UE with a satisfactory bearer, the UE changes from the serving cell to another through handover. In order to determine the most suitable time to perform a handover, the UE measures the signal strength of the serving as well as the neighbor cells in a regular basis. In E-UTRAN the eNBs can perform direct handover via the direct interface (X2) between eNBs. Very often though, the X2 interface is not available between (H)eNBs. In such a case, the eNB can initiate a handover involving signaling via the core network (S1- based handover).

A well-established HO policy in LTE is the «Strongest cell handover decision policy». The strongest cell, in terms of the received power can be based either on RSRP or RSRQ metric. The trigger of the mechanism can be activated by some pre-defined events: A2, A3 and A4. The A2 – A4 RSRQ algorithm is triggered when either A2 or A4 event is triggered. A2 event describes the situation when the serving cell’s RSRQ becomes worse than a pre-defined threshold, while A4 the situation when a neighbor cell’s RSRQ becomes better than another pre-defined threshold. The A3 – RSRP mechanism is triggered similarly, when the A3 event is triggered, i.e., when the RSRP of a neighbor cell becomes better that the RSRP of the serving cell.

In the following paragraphs, we discuss briefly the horizontal and vertical handover cases along with their main scenarios.

Handover to/from femto cells

Small cells are considered as one of the most promising solutions for macro-cellular network layouts. Small cells are low-power, short-range as a result, and relatively low-cost cellular access points that, although support fewer users than the macro cells, may significantly enhance the system capacity. Femto-cells usually feature self-configuration, self-optimization, self-healing, as well as advanced radio resource, interference and mobility management schemes. In a two-tier macro/femto cell environment, the following handover execution scenarios are identified:

Table 5: HO Execution Scenarios

HO scenario	Serving cell	Target cell	Access Control	HO Type	HO Execution Interface
1	eNB	eNB	Does not apply	Regular E-UTRAN	X2
2	HeNB	eNB	Does not apply	Outbound from HeNB	S1
3	(H)eNB	HeNB	Yes	Inbound to HeNB	S1
4	eNB	HeNB	No	Inbound to HeNB	S1
5	HeNB	HeNB	No	Inbound to HeNB	X2/S1

LIPA and SIPTO are two femtocell, -as well as Wi-Fi off-loading methods- for coping with traffic in the backhaul and core part of the network ([82]). The aim of these mechanisms is to avoid traversing the core part of the network, when the traffic for 2 UES is within the

local IP network, –a case in which direct connection would be sufficient -, or when the traffic originates from or terminates within the public Internet, a case in which the data can be offloaded either within the broadband backhaul network or at the femtocell. The LIPA function enables an IP capable UE connected via a HeNB to access other IP capable entities in the same residential/enterprise IP network without the user plane traversing the mobile operator's network except HeNB subsystem ([82]). The SIPTO function enables an operator to offload certain types of traffic at a network node close to that UE's point of attachment to the access network.

LTE interworking with GSM, UMTS, HSPA and non-3GPP access technologies

One of the most crucial aspects of the mobility management in the LTE networks is interworking with existing access networks while retaining IP connectivity. The EPC architecture allows for “session continuity”, that is that an IP connectivity session, which is established over any of the allowed access networks actually will survive movements between the different access networks due to loss of radio coverage. 3GPP has defined two different options for how to interconnect LTE and WCDMA/HSPA or GSM/GPRS. Similarly, with the MME for LTE, in WCDMA/HSPA/GSM/GPRS access technologies, the Serving GPRS Support Node (SGSN) is the one that serves the UE (maintains IP continuity, etc.) upon attachment with the network.

Besides the interworking with the afore-mentioned 3GPP access technologies, the vertical handover mechanism provides connectivity between LTE and non-3GPP networks, such as Wi-Fi. Although cellular and non-3GPP technologies originated from two fundamentally different objectives, -on the one side voice communication, and on the other side wireless data communication respectively-, the last years a need for interworking between the two has emerged. Two initiatives are currently being developed; on the one hand, Hotspot 2.0 by the Wi-Fi Alliance, and on the other, I-WLAN (Interworking WLAN) by the 3GPP ([83]). Hotspot 2.0 focuses on network discovery, authentication and roaming between networks. I-WLAN initiative aims to integrate WLAN technology into the core mobile network of the EPC architecture. Already, since the LTE release 10 and thereafter, an extended functionality to the -per UE- handover has started being defined, described in the MAPCON (Multi-Access Packet Data Network Connectivity) and IFOM (IP Flow Mobility) work items. The mobility management is gradually shifting from a single UE to the PDN connections as well as the IP flows, associated with the particular UE. In addition to the afore-mentioned technologies, the Access Network Discovery and Selection Function (ANDSF) ([84]) is an EPC entity, the role of which is to assist a UE to discover non-3GPP RATs, such as Wi-Fi or WIMAX.

3.4.2 Related papers

After presenting the standardized procedures by 3GPP in the previous section, in this section, we provide an analysis of the state of the art with regard to the handover mechanisms, which can be found in the literature. There has been a lot of effort into further optimizing the standardized mechanisms, and plenty of proposals and algorithmic solutions into further improving the policies, which lead the network and/or the UE to the handover procedure. In this section, a detailed categorization of these mechanisms is presented. More than half of the proposed mechanisms ([85], [88]-[91],[102]-[107],[111],[112]) rely on RSS/RSRP/RSRQ-based criteria. The signal strength and the signal quality, although often not adequate, are used as the first criterion before proceeding to more complicated handover decision algorithms. Furthermore, diverse context information sources are suggested from almost the schemes, such as the UE mobility speed and mobility patters ([85],[87],[92]-[94],[102],[109],[111],[112]), network or traffic-related parameters ([85],[93],[98]-[106],[108],[111]-[114]), or power-related

parameters, such as the UE power class or battery ([85],[95],[97],[102],[104]-[106],[108],[113]). As it can be inferred, a scheme usually takes advantage of more than one of the afore-mentioned input parameter types, attempting this way to optimise the final handover decision.

In [85] the authors attempt to optimize the classic handover trigger mechanism in LTE. What is claimed is that in the existing mechanism the handover trigger depends on the measurement report trigger timing in the UE, as well as the delay of the measurement report. The proposed mechanism is based on separate signaling mechanisms to identify the target cell and to trigger the handover. In addition, Channel Quality Indication (CQI) thresholds are taken into account as well –apart from the classic RSRQ metric-, in order to optimize the handover trigger parameters.

In [87] the authors propose an advanced UE history information based handover prediction mechanism. In 3GPP Release 8 the UE history information recorded by the eNB includes the Cell ID, as well as the time the UE stayed in a cell. What is described in this work is including Region-Domain Time-Domain, as well as Time-To-Trigger parameters into the UE history, in order to reduce the handover and Ping-Pong handover rates. Via simulation results they show that the proposed scheme on the one hand reduces the handover failure rate, while at the same time reduces the frequency of handover phenomena.

The survey by Xenakis et al. ([85]) presents an overview of the main HO decision criteria in the current literature and presents a classification of existing HO decision algorithms for femto-cells. A significant number of efforts, according to the authors, emphasize on the received power, prioritizing the femto-cell access and without taking into account context-related parameters like the profile of the user, the base station load or the speed of the UE. As the authors describe, in those schemes the final HO decision outcome is based either on the RSRP measurement, the Received Interference Power (RIP) metric, or the RSRQ measurement performed at the UE ([88]-[91]). According to the same survey, other efforts comprise received power – based decisions linked with the UE mobility information (speed, past patterns, etc.) in order to reduce the unnecessary handover attempts for high-speed users ([92], [93], [94]). Key energy-efficient criteria are the UE battery power ([95]), the mean UE transmit power ([96]), and the UE power consumption ([97]). The type of the traffic is one of the widely used parameters as well ([93], [98]-[101]). Some proposals, finally, combine multiple criteria (battery lifetime, traffic type, cell load, RSS, speed, etc.) relying on cost functions to produce the final outcome for the handover decision. Summarizing, the majority of the mechanisms listed focus on very specific criteria in order to take the final HO decision, without taking into account wider context-related information regarding either the UE or the network side. None of them takes into account the preferences of the user or attempts to relate the mobility pattern to the traffic type thus, failing to exploit user past information or build in such a way a complete user profile, towards which the decision for a HO could be optimized.

In addition to the previous survey, which discusses the macro to femto-cell HO decision policy in principle, the authors in another survey ([102]), present an overview of the vertical handover mechanisms available in literature, with Wi-Fi, WiMAX and UMTS are the main heterogeneous RATs. To begin, what is described in the survey, is an initial categorization of the information parameters of the VHO processes found in the literature into layers: application (e.g., user preferences, speed), transport (e.g. network load), network (e.g. network configuration, topology, routing information), data-link (e.g. link status) and physical (i.e. available access media). From the network perspective the ones highlighted are: latency, coverage, RSS, RTT, number of retransmissions, Bit Error Rate (BER), SINR, packet loss, throughput, bandwidth, network jitter and number of connected users. From the UE perspective the parameters, which are presented are user monetary

budget, preferred network (user choice), location, movement (change of direction), velocity, technologies available in the device, as well as battery consumption. Many of the proposed mechanisms that this survey presents attempt to create an overall context-aware mechanism, by combining several of the afore-mentioned parameters for the VHO decision outcome. Some of them do attempt to take into account user-oriented information; however, none of them, actually links this user-oriented information to the actual traffic generated by the user, as well as network metrics, in order to properly build the user's profile.

Having an overall look in the afore-presented survey on VHO handover mechanisms, it is made clear that several efforts attempt to combine various parameters from both the network as well as the UE side. None of them seems to take into account the afore-mentioned types of past information in order to address the issue of creating a proper user-profile based on the mobility of the UE, the traffic type that is generated and the user preferences though.

Several other existing surveys attempt to present a unifying perspective with regard to HO mechanisms (vertical or not). The authors in [103] deal with the network selection concept as a perspective approach to the always best connected and served paradigm in heterogeneous wireless environment. The HO decision criteria that may be considered in network selection process can be subjective or objective with minimizing or maximizing nature. From the origin point of view, they classify them in four categories: network-related criteria, terminal-related, service-related (e.g., QoS) and finally, user-related (e.g., user preferences). In addition, in [104]-[106], several efforts are described which aim to improve the selection mechanisms, which support heterogeneous RATs. In principle, all mechanisms combine several parameters like RSS, bandwidth, mobility, power consumption of the UE, security, monetary cost and user preferences.

Beyond the overview surveys presented earlier, we also discuss some latest research proposals. In [107], the authors deal with handover decision based on predicted received signal strength (RSS) of networks and dwell time. The dwell time is adapted in order to reduce the ping-pong effect or unnecessary handovers. In [108], a different idea is presented: two "conflicting" objectives are being dealt with in the proposed handover mechanism, i.e. the load balancing and the energy consumption. The user part's objective is the decrease of the energy consumption, while the operator's part benefits from the load balancing among the different cells. The presented simulations in [108] show that applying the hybrid proposed mechanism is beneficial for both the operator and the customers. Adoption Direction Prediction and Adaptive Time To Trigger (TTT) have also been proposed plenty of times to minimize the number of unnecessary handovers ([109]). The proposed mechanism however, although it takes into account the UE's previous location and estimates the direction of the UE, acting in a proactive way, it does not address at all the relation between the actual traffic type and this mobility information, in order to correlate them and create a profile for the particular user.

With regard to the processing of the input parameters after their aggregation, in the literature many authors have proposed Fuzzy Logic Inference Systems ([110]) in order to assess the inputs and generate an HO decision outcome ([111]-[114],). Fuzzy Logic (FL) is a tool for handling multi-variable problems, where a joint correlation analysis of several inputs is required. Indicatively, the authors in [111] propose a scheme, which takes into consideration the actual RSS, as well a predicted RSS, and they combine it with the speed of the UE in order to determine if a handover should be made or not. Moreover, they estimate the suitability of a RAT for handover, taking as input the current RSS, the estimated RSS, as well as the available bandwidth. In [112], the authors describe a multi-criteria RAT selection mechanism for 5G networks. Diverse context input parameters are taken into account, using FL, for the handover decision such as the UE's mobility, the

load of the base stations, as well as the type of service related to the handover. In [113], the authors propose a mechanism, which takes as input the available bandwidth, the delay, the charging information and the power consumption of the UE in order to calculate the most suitable RAT. In addition, they use GPS, in order to adapt the monitoring rate of the afore-mentioned values. In [114], once more, the bandwidth, jitter, Bit Error Rate (BER) and delay are used as inputs into a FL component to decide whether a vertical handover should take place or not.

Summarizing, it is becoming obvious that a large number of efforts can be found in the literature with many different approaches and different focus. Most of the proposals are taking into account RSS and interference-related metrics (RSRP, RSRQ, RIP, etc.) as inputs for the final decision. Moreover, the majority of the afore-described mechanisms are using some network-oriented metrics for evaluating the HO decision, i.e. latency, bandwidth, jitter, coverage, number of connected users etc. A considerable amount of the existing effort is focusing on the energy efficiency, trying to minimize for both the UE and the network sides the consumption. UE mobility information is also common (i.e. location, direction, velocity) in the literature. Last but not least, the user preferences are often taken into account, usually in the form of the budget and price preferences, as well as, type of RAT user preferences. Table 6, which follows, provides an overview of all the afore-described literature efforts, aggregating all the discussed parameters. According to the above reporting, and to the best of our knowledge, no proposal exists in the literature, which properly addresses the user profile creation based on past measurements related to the mobility of the UE, specific data traffic patterns (with varying requirements according to the specific service) and the preferences of the user with regard to cost and RAT technology.

Table 6: Handover schemes state of the art – Decision parameters overview

Papers	Signal Strength - Interference		Speed/ location			Network Parameters – Traffic Type			User & Device related info	Energy-related		
	RSS/RSRP/ RSRQ	Channel Quality	UE speed	UE residence time in the	UE location – mobility	Bandwidth/ Cell load/ Cell Capacity	Latency, Throughput	Traffic Type	Preferred network, device capabilities	UE power class	UE battery power	UE transmit power
[85]	X	X										
[87]			X	X								
[88]	X											
[89]	X											
[90]	X											
[91]	X											
[92]			X									
[93]			X					X				
[94]					X							
[95]											X	
[96]												X
[97]										X		
[98]								X				
[99]								X				
[100]								X				
[101]								X				
[102]	X	X	X		X	X	X		X		X	X
[103]	X						X	X	X			
[104]	X					X			X	X		
[105]	X					X			X	X		
[106]	X					X			X	X		
[107]	X											
[108]						X				X		X
[109]			X		X							
[111]	X		X			X						
[112]	X		X		X	X		X				
[113]						X	X		X	X		
[114]							X					

3.4.3 Related patents

In addition to the analysis that was presented earlier, based on the research and the proposed solutions and mechanisms found in literature, a significant number of the patents have been claimed from the researchers ([115]-[128]). A considerable number of patents claim methods based on the processing of network parameters, such as the base station load, the network throughput, the available bandwidth, etc. ([115],[117],[124],[125],[127]-[128]). Some of the efforts take into account application-related information ([115],[123],[124]), while this may be combined in some cases with location patterns of the user and prediction techniques based on historical user location information ([115],[123],[124],[128]). Finally, user related parameter-based schemes are claimed, such as user preferences, RAT selection per application type, cost preferences, etc. ([124],[127],[128]). Further insights with regard to these published patents is presented below.

In [115], the claimed application relates to mobility management in mobile networks based on context information and making a decision on a network service. The inventors by context-related information they are referring to one or more of the following: user profile, user history, network location and/or network topology, network capabilities, network services, charging models, potential next access points, location information, location prediction. According to the patent's description, however, no correlation between these inputs is made towards building on overall user profile that will be taken into account for the management of the user's mobility.

In [116], the described application includes a method for enhanced handover procedure. This method comprises transmitting an indication from the HeNB GW to a source HeNB connected to the HeNB GW, in response to receiving a UE Context Release Request message with an explicit GW Context Release Indication. The indication includes information on (a) whether the UE context stored in the HeNB GW has been released or not, (b) whether the target HeNB is connected or not to the same HeNB GW with the source HeNB and (c) whether or not the HeNB has to indicate to the HeNB GW that a subsequent handover from the source to the target HeNB has been performed.

The patent in [117] relates to a handover method for redirecting an on-going communication of a connected UE from a serving cell to a target cell. The method comprises receiving (a) a cell attenuation measurement, (b) a supplemental hysteresis parameter, which load depends on the load of the target cell and (c) a threshold parameter.

In [118], the authors describe scenarios, in which a UE is handling data traffic requiring a high level of QoS, while also handling traffic for which best effort delivery suffices. In [119] the patent describes a UE capable of storing identity information associated with a source cell, and using this information in accessing other target cells subsequent to radio link failure so as to facilitate access to context information of the UE. [120] describes a system and a method for enhancing using Mobility Management (MM) and Session Management (SM) procedures in an SAE/LTE system. These procedures modify several MM and SM procedures by 3GPP for other access systems (e.g. UTRAN). With respect to the signaling, the present invention re-uses the existing information element definitions for each parameter included in a message.

[121] claims a method for supporting handoff from GPRS/GERAN to LTE E-UTRAN, relating to a method for use in LTE MME comprising: Receiving a relocation request message from an SGSN serving a wireless UE using GERAN technology, communicating a handover request message to a eNB and executing the handover notifying the previous RAT, as well as providing LTE services to the UE. In [122], the patent describes an off-loading method from a cellular network to WLAN network and vice versa. The method

comprises the steps of automatically detecting call handover threshold for a UE from a specific RAT to another; selecting said second call device from a set of previously defined target handover devices; and establishing a connection to the target RAT.

[123] proactively deploys a handover decision mechanism in relation to a handover and in view of the operational context of the mobile device. The method comprises as well determining at least one new access point for the mobile device using the mechanism. The context is defined as the current location, the movement prediction, the UE speed and the profile of applications being executed in the mobile device before or at the time of pro-active deployment of the handover decision mechanism.

In [124], a system and method that facilitates optimizing handover is being claimed. A rules engine can create a user-defined rule based upon at least one of a resource requirement for each device application or a user input related to a quality of service (QoS). In [125] a method is claimed for performing a handover in between heterogeneous access networks. The method comprises receiving from an access network information transfer manager being common to the plurality of heterogeneous access networks, available access network information containing a list of available access networks in the vicinity of the mobile terminal in form of a context. The system claimed in [126] includes a mobility manager operable to receive link quality evaluation messages (MNE) from the mobile nodes providing an indication of a currently available link quality from an access network with which the mobile node is currently affiliated.

In [127] a method for deciding a handover of a mobile terminal based on context information is claimed. The patent comprises evaluating selection priorities of network interfaces for each transmission path for each application program based on the context information and obtaining a handover decision value for selecting a best network interface from among the networks by applying a cost function to the evaluated selection priority. Finally, in [128] a priority for selecting a cell on request for the handover traffic service is determined by considering user preferences and status of neighboring cells, and an optimized cell is determined according to the determined priority. The priority is determined by considering user preferences for traffic, status of cells having traffic, battery status and velocity of mobile terminals, and traffic priority and emergency.

A summary of the above-described HO patents is provided in the table that follows (Table 7). The afore-presented analysis gave an overview of the current state of the existing patents, each one focusing on several perspectives either for enhancing the handover procedure itself, or in order to obtain a more holistic knowledge of the network environment and –as a result- optimize the handover decision. Some emphasize on context-related information collection, while others focus on several other technical enhancements. An initial observation is that some patents focus on various technical enhancements of the handover procedure without focusing on network-related information, or user-related context-information. Several claimed patents are based on network related information such as network load, bandwidth, delay etc. A smaller number of these -besides network-related context- take into account user-related information such as application info, prediction of location and movement, as well as user preferences such as cost or RAT type per application preferences. None of them, however, seems to actually correlate user history-based information to applications, past location and mobility patterns, device characteristics and capabilities into finally building generic user profiles, which will be used in order to realize traffic engineering from a more holistic network point of view.

Table 7: Handover patents overview

Patent reference	Network-related (load, bandwidth etc.)	Application related information	Location info (location, speed)	Prediction of location/ movement	User preferences	UE info (battery status, etc.)	History-based UE Profile	Other technical enhancement
[115]	X	X	X	X			X	
[116]								X
[117]	X							X
[118]								X
[119]								X
[120]								X
[121]								X
[122]					X			
[123]		X	X	X				
[124]	X	X	X	X	X	X		
[125]	X							
[126]								X
[127]	X				X	X		
[128]	X		X		X	X		

3.5 Latest efforts addressing traffic steering and RAT selection

Lately, numerous novel efforts found in journal papers, which address the main RRM mechanisms, traffic steering and RAT selection procedures for future networks have been published ([142]-[152]). Moreover, the last three papers ([150]- [152]) take into account – besides other parameters- ANSDF features of the latest 3GPP release.

In [142], the authors focus on the handover delay challenges, from the handover security and user authentication perspective. Ultra Dense deployments may result in frequent handovers, which may subsequently introduce high delay overheads. They propose the Software Defined Networking (SDN) enabler as one of the most promising solutions; through its centralized control capability, user-dependent security context may be exchanged between related access points and enable delay-constrained 5G communications. The context is shared between nodes and APs based on UE path prediction. A redesign on the “intra-macro cell” handover procedure is described in [143], focusing on the control/user plane split HetNets of future systems; the handover optimization is realized by predicting the received signal quality of the UE, triggering as a result the handover decision in a more efficient way. The authors focus on the challenges of the handover between macro cell at high speed scenarios (railway, highway, etc.). The respective evaluation shows that by predicting the forthcoming UE measurement reports, the handover execution takes place in advance and the handover performance is enhanced.

Similarly, in [144], the authors also focus on the -control and user plane separated- future HetNets and more specifically, on the signaling latency reduction in cases of macro cell base station fail-over periods. The proposed solution is based on a small cell controller scheme for controlling and managing small cells boundaries in a clustered fashion, during the corresponding macro cell’s fail-over period. The evaluation of the proposed scheme

on the UE side demonstrates reduced signaling latency, particularly for high user velocities; however, at the same time, the data delivery latency increases comparing to the legacy scheme; the authors conclude that the application of the proposed scheme can be selected on the specific signaling and data delivery latency requirements of each use case.

The RAT selection and handover procedure have been also studied from lower levels' perspectives as well. The very recent work in [145] focuses on the high frequency bands above 6 GHz, which will provide considerably larger bandwidths than the legacy systems; the challenge that the authors identify relates to the modifications that need to take place in the design of some key functions, such as the handover in order to support future deployments. They propose a novel frame structure, flexible and scalable to support various numbers of beams/antennas, users, or traffic conditions. The evaluation that was conducted involved static, as well as high velocity UEs; the authors conclude that the proposed enabler succeeds at satisfying all the throughput and delay requirements of the forthcoming 5G and beyond use cases.

Handover management in ultra dense heterogeneous small cell networks is studied in [146], focusing on the cell edge users. The authors describe an architecture comprising a cloud radio access network (C-RAN), as well as base band unit (BBU) pools, in which resource management and control capabilities are co-located, such as handover decision function and admission control. The proposed handover is realized between the BBU pools. The evaluation of the proposed scheme showed that the capacity of the small cells is increased, without increasing however the QoS of the users as well. In [147], the authors outline the main challenges that come with the UDNs. Among numerous challenges, such interference mitigation, backhaul issues and energy consumption, the authors tackle the mobility and handover challenge as well. Among the enabling technologies they propose is cell and receiver virtualization, self-backhaul solutions and user-centric control of user information to minimize signaling.

The challenges posed by the UDNs are discussed also in [148]. The authors present an overview of the existing solutions related to control and data plane separation architecture (SARC), which they consider as one of they key enablers of the UDN use case, while they focus on the coordinated multipoint (CoMP) and Device-to-Device (D2D) communications. According to the authors, SARC can optimize considerably the handover mechanism and minimize the signaling overheads that were imposed until now: for example, users with no active data sessions will not have to be handed over. Similarly, in [149], the authors study the 5G UDNs challenges and propose a novel architecture that absorbs the Machine Type Communication (MTC) high and unpredictable traffic via home eNBs, allowing them to significantly reduce congestion and overloading of radio access and core networks. The main goals that the authors address are a stronger separation between MTC and Human Type Communication (HTC), closed access femto cells , -which will be only available to those machines that belong to a given closed subscribed group-, as well as coverage extension. In relation to the handover mechanism, the primary claim of this work relates to the X2 interface, serving a low-latency interface to exchange data traffic for time critical events of MTC traffic. Via extensive evaluation scenarios involving MTC UDNs, the authors demonstrate considerable gains in terms of latency in the handover procedure, energy consumption, capacity and scalability.

In relation to the ANDSF network entity, which will play a major role in the future Dense HetNets –as discussed earlier-, we must highlight the fact that the only literature proposals attempting to address the 5G RAT selection topic taking into account the ANSDF features of the latest 3GPP release are [150], [151] and [152]. In [150] the authors provide detailed insights with regard to the means that the context information is acquired and by describing how ANDSF is utilized for retrieving the desired information items, i.e.,

the BS transmitted power, the cell traffic load and the user's spectral efficiency. Nevertheless, the authors do not address the issue of the frequent handover process trigger as they solely rely on the A2 handover event (RSRQ threshold), while in addition no reference is being made to the velocity of the UE, an essential parameter need to be taken into account for also avoiding unnecessary handover triggers and ping-pong effects. Finally, no discrimination is being made depending on the type of the traffic flow, implying that all traffic flows –no matter their specific QoS requirements- are handled the same way. On the other hand, in [151], the authors choose to use the ANDSF to facilitate the discovery of non-3GPP access networks; however, the context information that may be residing in ANDSF is not claimed to be taken into account for the optimization of the handover procedure, as the authors select to follow the existing handover techniques. Finally, in [152] the authors propose an energy-efficient handover mechanism, based on the ANDSF, which attempts to minimize the overall power consumption, maintaining a minimum QoS for the active sessions.

In most cases solutions target either handovers for macro-femto cells or vertical handovers among different RATs as separate approaches. In this thesis, from now on we attempt to overcome such discriminations and approach the problem from a unified perspective, applying the conclusions, which result, to all three RAT selection mechanisms described earlier.

4. COMPASS: CONTEXT-AWARE, USER-DRIVEN, NETWORK-CONTROLLED RAT SELECTION FOR 5G NETWORKS

This section introduces COmpAsS, a real-time context-monitoring, UE-oriented RAT selection and traffic steering/switching mechanism for UDN Heterogeneous Networks towards 5G. The section initially provides an overview of the mechanism; then it discusses the network architecture perspective taking into account the latest 3GPP developments, discussing also in detail all the extensions needed in the network interfaces that enable the exchange of the required context information among the respective network entities; afterwards, the proposed scheme is described in depth, both from the algorithmic, as well as the Fuzzy Logic modeling perspective; the last parts of this section comprise an evaluation of the scheme: on the hand the signaling overhead perspective is evaluated; on the other, extensive simulation results are presented in a complex and realistic 5G use case, illustrating the clear advantages of the proposed approach in terms of key QoS metrics, i.e. the user-experienced throughput and delay, both in the uplink and the downlink, proving thus the validity and viability of COmpAsS for context-based RAT selection.

4.1 Introduction

The evolution of the Access Network Discovery and Selection Function for the core part of the cellular network, as well as the Hotspot 2.0 approach –as presented earlier in this thesis-, are currently being subject to thorough discussions and studies and are expected to facilitate a seamless 3GPP-Wi-Fi interworking. Furthermore, as also illustrated in the previous section in a comprehensive analysis of the state of the art, during the past years, several RAT selection schemes have been proposed, however, these none of these schemes do not take into consideration the opportunities offered by the aforementioned new standardized approaches, i.e. ANDSF and Hotspot 2.0.

Optimizing the traffic steering strategy among dense heterogeneous network scenarios is one of the compelling challenges in the latest releases to be addressed, as well. Although, already before Rel.12, the Wi-Fi interworking was supported at the core-network level (via IP-session continuity supported by GTP, MIP and PMIP protocols, the Wi-Fi and Public Land Mobile Network (PLMN) selection mechanisms were not satisfying enough to be deployed, and the traffic steering mechanisms via ANDSF were not sufficient. What was missing was a mechanism to determine the access network to use for each given IP flow. In the current release, ANDSF has been enhanced introducing the Inter-System and Inter-APN Routing Policy (ISRP and IARP) rules ([175]), which enable the UE to determine the access network to route its active flows, by taking into account besides the Received Signal Strength Indicator (RSSI) levels, the Wi-Fi base station load as well. Our mechanism that will be thoroughly described in the following sections, not only does utilize the ANDSF; it takes the aforementioned one step further, creating an even more holistic picture of the network conditions of the candidate RATs in a lightweight and efficient manner.

Towards this direction of addressing the afore discussed challenges and following the foreseen advancements on the road to 5G HetNets, this thesis proposes a novel mechanism, which follows closely the latest 3GPP directions and guidelines and attempts to cover the aforementioned gaps. More specifically:

a) This work concentrates on the context acquisition process: A comprehensive analysis on the network sources, respective interfaces and context information item types is made. In addition, an analytical approach is presented, which provides detailed insights on the information items, which are used, along with signaling overhead required to aggregate them. To the best of our knowledge,

there is no previous work, which attempts to quantify the signaling overhead of the proposed context-based mechanism.

b) The mechanism is extensively evaluated from the performance perspective: The proposed scheme is a lightweight module, the core of which is based on a fuzzy-logic inference system. The validity of our proposal is evaluated via numerous simulation scenarios and diverse traffic flow types, in a realistic 5G configuration and set-up, comprising an ultra-dense heterogeneous network environment.

c) The novelty of this work is further reinforced by the fact that the proposed scheme is based solely on assumptions in line with the latest 3GPP standardization efforts in terms of context information acquisition, attempting this way to highlight the realistic and viable aspect of the solution for next generation wireless networks. To the best of the writer's knowledge, no research proposal has attempted to limit its assumptions totally in line with the standardization guidelines; on the contrary, the vast majority of solutions make numerous assumptions, which often lead non-realistic proposals.

COmpAsS has been already validated in two preliminary works ([176],[177]), as well as a third extensive analysis ([178]), which evaluates the proposed scheme in a diverse set of scenarios, traffic types and mobility scenarios.

During the early evaluation of this scheme in the aforementioned works, an initial, basic rules set was evaluated, upon which the fuzzy logic scheme was based, while also the first simulations to initially assess the validity of the proposal were performed. In the final, evolved COmpAsS version, the algorithm of the scheme has been carefully enriched from numerous perspectives, which will be also explained in detail in one of the next sections: the signaling cost perspective of the mechanism has been thoroughly investigated; there is always a crucial trade-off for context-based scheme, and this is the first time an assessment is being made relating taking also in account the current (3GPP's) available information items as well; from the algorithmic perspective, the mechanism of the triggering events has been optimized: an optimized *Threshold* and *Hysteresis* mechanism (see next sections) is provided that optimizes its functionality from the energy consumption, as well as network signaling perspective; furthermore, the rules set, which is applied on the Fuzzy Inference System has been progressively fine-tuned and optimized after numerous simulations and feedback loops assessing pre-defined Key Performance Indicators (KPIs); next, the environment of the simulation has become more realistic and sophisticated, having also added a building propagation model, as well as a shadowing loss model, -which were previously missing-; last but not least, in this comprehensive round of simulations, multiple traffic types to the UEs (VoIP, FTP, etc.) are assigned, studying the behavior of the scheme and the fine-tuned rules set for completely diverse different types of IP flows (and QoS requirements respectively).

As far as the core evaluation mechanism is concerned, i.e. the mechanism, which receives the input context information, processes, evaluates it and generates the handover decision information, diverse solutions have been proposed and used. For the proposed solution, a Fuzzy Logic (FL) model has been chosen in order to support the decision making process. Several authors have proposed FL Inference Systems. Indicatively, Xia et al. [179] propose a scheme taking into consideration the actual RSS, as well a predicted RSS, and they combine it with the speed of the UE in order to determine if a handover should be made or not. Moreover, they estimate the suitability of a RAT for handover, taking as input the current RSS, the estimated RSS, as well as the available bandwidth. In [180], FL is also used for estimating the output suitability of a network based on the inputs of the environment (bandwidth, delay, charging, power

consumption). In addition, Ma and Liao use GPS, in order to adapt the monitoring rate of the aforementioned values.

4.2 Overview of the proposed solution

The proposed RAT selection mechanism that is presented in this work aims at enabling the UEs to identify in an intelligent way the most suitable RAT to associate with in a specific urban area, where a cellular operator (with deployed macro, pico or femto cells) co-exists with Wireless Internet Service Providers (WISPs) with whom it has roaming agreements. The mechanism is applicable for users of 5G smartphones that support a number of RATs.

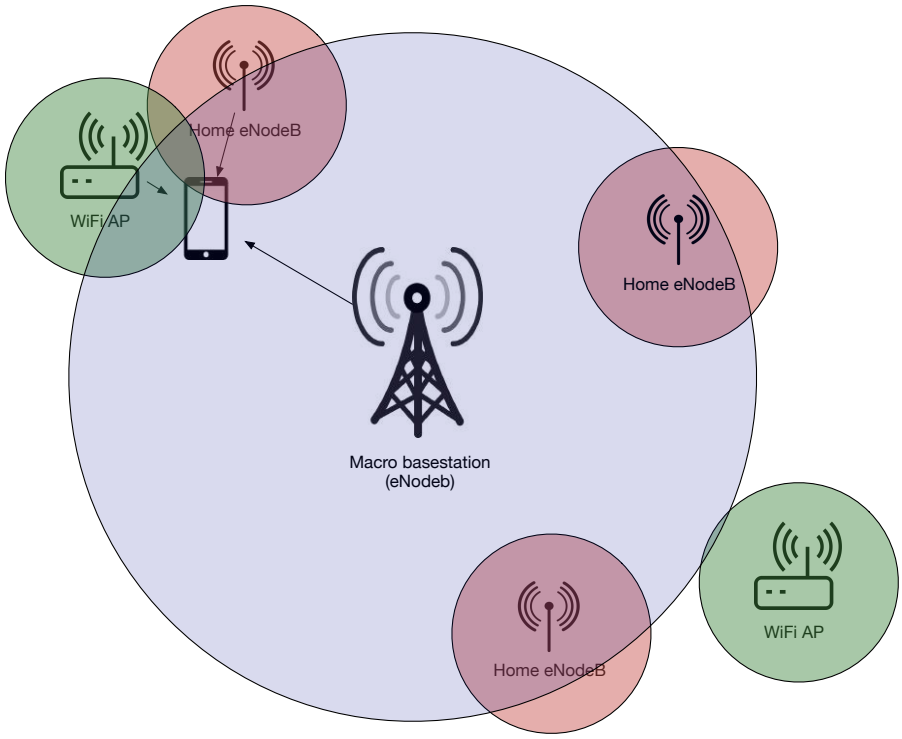


Figure 9: Context-based RAT selection by COmpAsS in a Heterogeneous 5G Environment

The framework is using pre-defined, customizable and fine-tuned rules for all the possible combinations of the different aforementioned scheme's inputs. The rules that are applied are policies, based on objective network parameters, KPIs and general principles, derived from the state of the art of the domain, as it was presented in the previous section. More specifically, according to these rules/policies, a RAT, which is characterized by low (backhaul) load and high RSS/RSRQ, is advantageous for the UE choice. In addition, the higher the sensitivity to latencies (traffic flow type input type), the higher impact the mobility metric has on the Suitability; high mobility UEs are preferably placed in larger cells to avoid unnecessary handovers and/or ping-pong effects. Using Fuzzy Logic Controllers (FLC) each UE evaluates the available RATs and identifies the most suitable one, which optimizes the Quality of Service (in terms of pre-defined KPIs) for each application (or type of traffic); afterwards it performs a session initiation or a per flow-handover using existing 3GPP mechanism described in the introductory section. The KPI that is utilized to describe the selection prioritization among heterogeneous cells and access technologies is denoted as *Suitability* in our scheme (Figure 10).

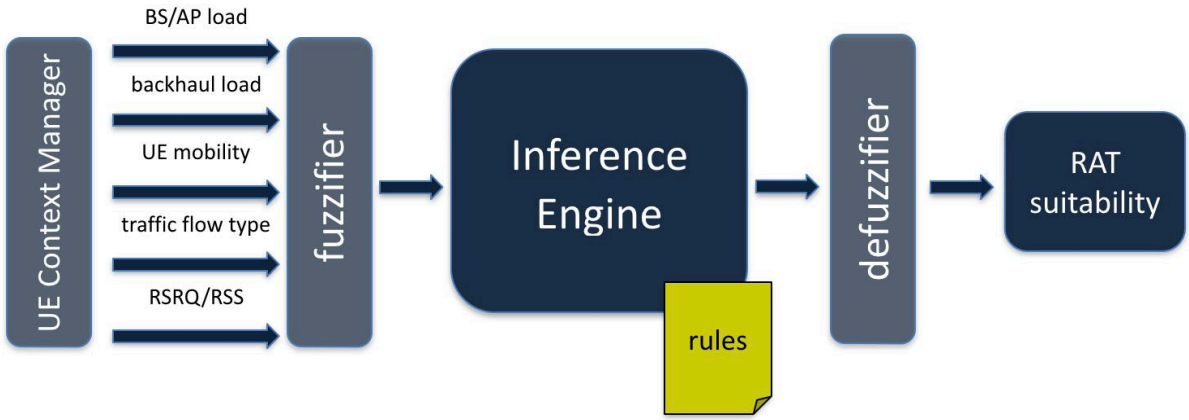


Figure 10: Fuzzy Logic Controller for the extraction of the RAT Suitability metric

In the proposed scheme, the UEs collect information from a local instance of ANDSF (Local ANDSF - L-ANDSF) as suggested in [15] about the policies of the operator for accessing Wi-Fis in the area, as well as information from Hotspot 2.0 protocols to evaluate the status of Wi-Fi APs (e.g., number of users associated to the AP, the load of the backhaul link of the AP etc.). We extend this concept by assuming that the L-ANDSF may contain similar information (i.e., number of associated users and load of the network link) for every (H)eNB in the area. This requires appropriate logical interfaces from the (H)eNBs to the ANDSF. These nodes can update L-ANDSF in a coarse manner (e.g., Load may be characterised as “Low”, “Medium” or “High”) only when thresholds are violated so as to minimize the required signaling exchange. The mobility pattern of a UE can be monitored either from the accelerometer of the smartphone, or as 3GPP suggests, from the number of the cell relocation actions that have been executed. Finally, the categorization of applications based on their sensitivity to latency can be extracted by a UE connection manager, from the well-known port numbers of the applications.

All in all, the *Suitability* metric relates separately to each one of the active UE’s traffic flows; in other words, for each active traffic flow F and for each available RAT R there is a different value, resulting thus, in N_F x N_R overall values, where N_F and N_R are the number of the flows or the available RATs respectively (see example in Table 8 for N_F =4 and N_R =5.).

Table 8: NF x NR Suitability Calculation example for a UE with 4 active IP flows

UE active flow #	RAT Suitability list
1: browsing (downlink)	*eNB3, WLAN SSID1, eNB2, eNB1, WLAN SSID2*
2: VoIP (uplink)	*WLAN SSID1, WLAN SSID2, eNB3, eNB2, eNB1*
3: VoIP (downlink)	*WLAN SSID1, WLAN SSID2, eNB3, eNB2, eNB1*
4: background cloud syncing (uplink)	*eNB1, eNB2, eNB3, WLAN SSID2, WLAN SSID1*

The proposed scheme’s decision-making process selects for each one of these active flows the RAT, for which *Suitability* is maximized; afterwards, the UE makes a handover request to the respective (H)eNB or AP in order to transfer the flow to the optimal access technology. The process is running both on a pre-defined time interval basis, as well as upon pre-defined trigger events, which are described in detail in the following sub-section.

If for any reason, the handover to the highest-ranking RAT is not possible, the 2nd choice in the *Suitability* list is selected, etc.

4.3 The network architecture perspective

The mechanism –as already mentioned earlier- is user-oriented, i.e. deployed on the UE side, rather than one of the main LTE network entities. Nevertheless, its functionality is not completely independent as the architecture of the network is directly influenced in the sense that the context information, which is required to be aggregated by the UE, is available in specific network components. Moreover, the mechanism is directly associated with the network policies, in the sense that the UE-oriented decisions are being forwarded to the central decision-making entity in the network core, which will make the final assessment. As a result, some minor adaptations need to be realized in order to enable the required context information acquisition. In this subsection, we address the requirements in terms of the data sources, message types, as well as interfaces, which are required to support the proposed mechanism.

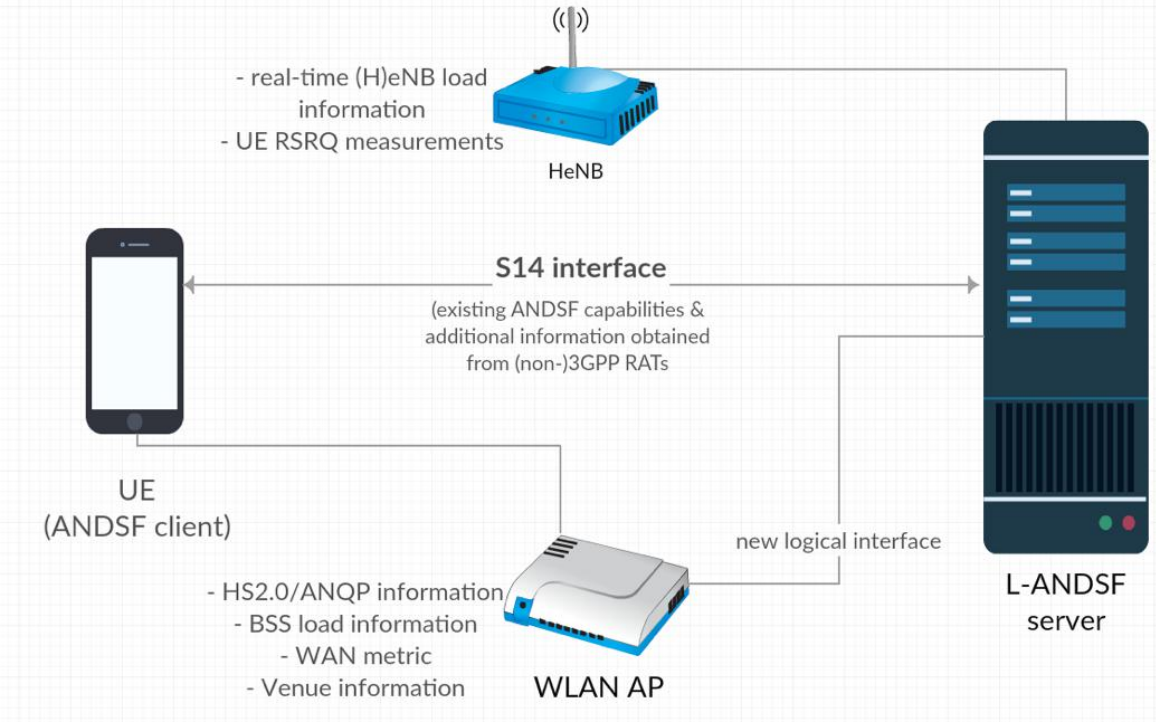


Figure 11: Main network entities including the proposed extensions of the Local-ANDSF (L-ANDSF) and respective interfaces

One of the core network entities, on which COmpAsS relies is the ANDSF. As described earlier in this paper, ANDSF is a cellular technology standard, which implements dynamic data offloading for the UEs in a structured way. However, the purpose of ANDSF is currently limited to provide the UE with policies with regard to access networks. Moreover, one of the most crucial aspects in relation to offloading and handover mechanisms, which is not contained ANDSF Management Object (MO), is real-time network conditions, such as the load of a Base Station. This type of information, as well as additional features, which are not provided by the ANDSF, may be provided by the Hotspot 2.0 standard described earlier, supported by the ANQP protocol.

On the contrary, ANDSF provides WLAN AP location information, supports UE location reporting, as well as may provide a list of preferred or restricted access networks, -features, which are not provided by Hotspot 2.0 -.

It becomes clear that ANDSF and Hotspot 2.0 could act in a supplementary way to maximize the available information to the UE, resulting in more efficient offloading

mechanisms. In this paper, we propose an enhanced version of the ANDSF server capable of:

a) collecting real-time load information regarding the available 3GPP access networks, based on a new logical interface (e.g., between the (H)eNB and the ANDSF entity). This information is evaluated in a coarse manner (i.e., low, medium, high).
b) supporting queries to Hotspot 2.0 enabled WLAN APs using the ANQP protocol
c) gathering information from the UE measurements regarding RSRQ measurements

In relation to the input parameters that have been taken into consideration for COmpAsS scheme, certain inputs are already available using the existing standards (thus no further assumptions are required), while for the rest, some additional assumptions regarding the applied protocol (message type, etc.), as well as the respective interfaces are required. As already mentioned briefly earlier, the UEs monitor the following contextual information items:

- the traffic load of the cellular base stations and/or Wi-Fi APs (in terms of available bandwidth)
- the backhaul load of the available access networks
- the mobility characteristics of the UE (speed, etc.)
- the type of the traffic flow (mapped to a specific sensitivity to latency for each flow type)
- the RSS (or RSRQ for 3GPP networks) of the available RATs/cells/APs.

More specifically:

- The *RSRQ value* is already part of the UE measurements report that is used in LTE for evaluating the quality of the signal of the neighbour base stations. Similarly, for Wi-Fi, RSS metric is already included in the existing IEEE 802.11 reporting metrics, even for the end-user devices.
- As indicated in [181] the *mobility state* of the UE (high-mobility state, medium, etc.) is considered and is sent via the system information broadcast from the serving cell.
- The *traffic flow type*, mapped to the respective sensitivity to latency for each application/service type executed by the UE: the different application categories and respective flow QoS requirements are extracted by the UE connection manager, from the well-established port numbers of the applications/services.
- The *traffic load* of the base stations and the *backhaul load* of the network: In the proposed mechanism, the UEs collect information from a local instance of ANDSF (Local ANDSF - L-ANDSF) about the policies of the operator for accessing Wi-Fis in the area, as well as information from Hotspot 2.0 protocols to evaluate the status of Wi-Fi APs (e.g., number of users associated to the AP, the load of the backhaul link of the AP etc.). The main functionality and role of the ANDSF has already been discussed. We extend this concept by assuming that local ANDSF entities (L-ANDSF) contain similar information (i.e., number of associated users, load of the network link, etc.) for every (H)eNB in a specific. This distributed model radically decreases the information exchange delays between the nodes in a limited area, comparing to a scenario, in which one central ANDSF entity of the operator serves thousands or millions of devices. This requires appropriate logical interfaces from the (H)eNBs to the ANDSF. Last but not least, the nodes update L-ANDSF in a coarse manner (e.g., Load is Low, Medium or High) only when thresholds are violated, so as to further minimize the signaling overhead in the network.

The obtained information is aggregated by a *Context Manager* entity, part of the proposed scheme, which resides inside the UE and processes the information in order to forward it to the Fuzzy Inference Engine, which is described in the following sub-sections.

4.4 Description of COmpAsS algorithm

Initially, and prior to proceeding in each one of the algorithm steps description, it is required at this point to provide some insights regarding two parameter types, which need to be pre-set prior to the algorithm deployment on the UE. Although the FL computational requirements are minimum, in order to further optimize the energy consumption of COmpAsS scheme inside the UE, as well as to minimize the unnecessary handovers, the algorithm is evaluating two types of parameters, namely:

- *a Suitability Threshold T (0%<T<100%):* the UE evaluates the current Suitability of its currently associated RAT/cell and compares it with a pre-defined algorithm parameter, namely the Threshold, above which the current RAT is considered as satisfactory for serving the UE requirements. For example, if the T=90%, no FL computation is performed (for the particular IP flow) if the associated (current) RAT's Suitability is above 90% (implying that the current UE's RAT is satisfactory enough to attempt any new handover).
- *a Suitability Hysteresis (Margin) value (0%<H<100%):* it describes the required advantage difference between the candidate cell's Suitability when compared to the current one's in order to consider it as preferred choice. Multiple Hysteresis values may be used for different target RATs, according to the planning of the network administrator, for example: if HMACRO=10% and HFEMTO=3%, examined RAT's Suitability must be at least 10% higher than the current RAT's, -if a neighbor RAT is a macro cell-, or at least 3% higher than the current RAT, -if neighbor RAT is a femto cell-, in order to trigger a handover towards the respective candidate RAT. The higher Hysteresis in the case of macro neighbor RAT may be chosen aiming to impel the handover to smaller RATs for offloading reasons. From a broader perspective, the customizable HMACRO and HFEMTO values as far as the Hysteresis is concerned provide the network administrator a wide range of options, being able to control the interworking balance between the macro and small cells, as well as dynamically route the offloaded traffic flows. Both the Suitability Threshold, as well as the Hysteresis parameter evaluation follow in the next algorithm steps.

The values of the *Threshold* and the *Hysteresis* may be configured according to the specific needs of a particular network environment by the network administrator before the mechanism is deployed on the UE. An extension of this feature that could also be accommodated in the future is the enablement of an automatic adaptation of the two control parameters, by defining the different possible "states" of the network and the respective *Threshold-Hysteresis* configuration for each one of these states. For example, for a denser, -in terms of network deployment- environment, the solution performs better for higher *Threshold* and *Hysteresis* values. It should be also noted that the network "state" sensing is already enabled via the available context information that is being aggregated by the UE.

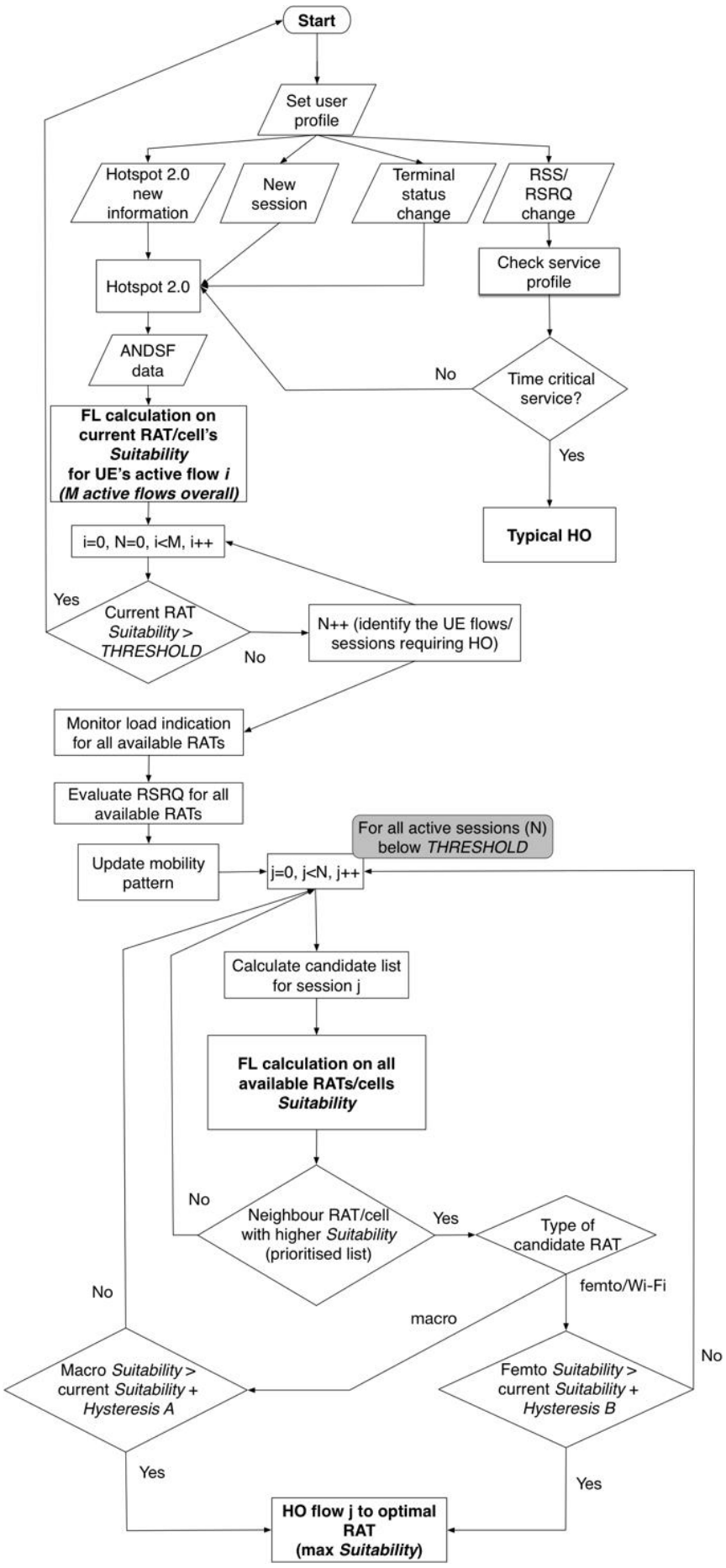


Figure 12: COmpAsS algorithm

The algorithm (Figure 12) is described thereinafter step by step: initially, the user defines either on a per session basis (e.g., HTTP traffic to be handled only by free Wi-Fi) or collectively (e.g., use always the RAT that minimize the energy consumption) his preferences (i.e., “user profile” in the algorithm flowchart). For the user profile generation, numerous solutions have been proposed, i.e. either manually or automatically using data analytics solutions; for our algorithm, a solution that suggests the profile creation in an automated manner is selected. The mechanism algorithm may be triggered only if there is at least one session active in the UE. Thus, as long as there is at least one session, pre-defined events trigger the algorithm initiation, i.e.: new information from Hotspot 2.0 is received (e.g., a Wi-Fi AP is now unloaded), a new session is initiated on the UE side, a significant change in the terminal status (e.g. battery level is falling below a certain threshold) or a significant change in the monitored RSS/RSRQ values is identified, etc.

By the time the process is triggered, and only if there is no time critical service to be served, (-in which case RSS/RSRQ has fallen below a threshold and a typical HO must urgently take place-), the UE proceeds to the information aggregation phase, which is being supported by Hotspot 2.0 and ANDSF servers, in relation to all the available neighbor RATs (3GPP or non-3GPP) and cell layers (macro, pico, femto cells, etc.), without any direct association with them. This information relates to the available 3GPP or non-3GPP cells’ and APs’ load, number of associated UEs, quality of received signal, etc. As already discussed in Section II, in 3GPP Rel-12 and beyond, the ANDSF enhancement creates a sufficient RAT context source. This is achieved by incorporating additional information items, better granularity for the existing for traffic steering conditions, as well as integrating information from Hotspot 2.0. This information is updated on the UE side, triggered on a per trigger-event basis. Having aggregated the updated context information from the aforementioned sources, the UE performs an updated calculation on the Suitability of the currently associated RAT/cell, i.e., whether it is over the pre-set *Threshold* value. If yes, then no further calculation or signaling is required and then algorithm reverts to the starting point.

However, if the *Suitability* of the currently associated RAT/cell is below the Threshold, the UE continues with the aggregation of the context information of all available RATs/cell layers (i.e. load monitoring, RSRQ indicators, mobility patterns). Then, for each one of the active sessions in the UE (i.e. all active flows, including applications download/upload, background services, etc.), the Suitability KPI is calculated for the available (candidate) RATs/cells. That results in a *Suitability* –based prioritized list for each one of the active sessions/flows. Starting from the top RAT/cell, and following a top-down approach throughout the priority list, an evaluation of the *Suitability* value takes place: for macro cells (LTE, GSM, etc.), the candidate RAT’s *Suitability* must be higher than the current’s RAT *Suitability* by *Hysteresis A* in order for it to be selected for session handover; for small cells respectively (i.e., LTE femto cell/Wi-Fi AP), the candidate RAT’s Suitability must be higher than the current RAT’s Suitability by *Hysteresis B*. As already discussed above, the network administrator/traffic engineer is able to control –without altering any other policies/rules- the offloading flow routes via dynamically adapting these two different *Hysteresis* values; by increasing A value comparing to B for example, the traffic engineer may target to induce session handovers to smaller RATs for offloading reasons.

The RAT/cell with the highest *Suitability* is being selected for starting the procedure of the handover; in case of a rejection the second in the list is being selected for initiating the same procedure, etc. It must be noted, that in case the priority RAT list is exhausted without satisfying the *Hysteresis* conditions (e.g., due to the fact that the *Hysteresis* value has been set too high), –and as a result, no handover has been decided and triggered-, the list is once more traversed without the *Hysteresis* values, in order to facilitate the handover realization.

4.5 Fuzzy Logic modeling of the solution

As already discussed earlier, we identify five input context parameter types, which are considered the core metrics for the evaluation of the *Suitability* of a candidate RAT/cell for handing over an active UE session. The acquired information is processed using a Fuzzy Logic Controller (FLC), which is the core decision-making mechanism of the proposed scheme. Fuzzy logic is an ideal tool for dealing with uncertainty cases, when the inputs are rough estimated values. Furthermore, FL handles the curse of dimensionality by using generic input and output states. Finally, the fuzzy logic is a tool for handling multi-variable problems, where a joint correlation analysis of several inputs is required.

Each FLC is being composed of the fuzzification component (fuzzifier), the inference system (FIS), and the defuzzification module. The fuzzifier undertakes the transformation (fuzzification) of the input values to the degree that these values belong to a specific state (e.g., low, high, etc.). Then, the FIS correlates the inputs and the outputs using simple "IF...THEN..." rules; each rule results to a certain degree for every output. These rules relate to the network administration policies and could be as a result related to traffic engineering rules. Thereinafter, the output degrees for all the rules of the inference phase are being aggregated. The output of the decision-making process, comes from the defuzzification procedure. This degree may be obtained using several defuzzification methods; the most popular is the centroid calculation, which returns the center of gravity of the degrees of the aggregated outputs.

For the purposed of this thesis, we have used the MATLAB software, along with the Fuzzy Logic Toolbox[4]. The proposed scheme uses three FLCs, each one for every RAT type, i.e., macro cells (LTE, GSM, etc.), femto cells and Wi-Fi APs. Every time that the algorithm is being triggered, all the available base stations and APs are being evaluated. The inputs of every fuzzy reasoner are the RSRQ/RSS, the service sensitivity to latency, the load of the (H)eNB or AP, the backhaul load of the corresponding (H)eNB/AP, as well as the UE mobility status. The modeling of the 3 FLCs inside MATLAB is illustrated in the following figures:

[4] MATLAB Fuzzy Logic Toolbox, https://www.mathworks.com/products/fuzzy-logic.html

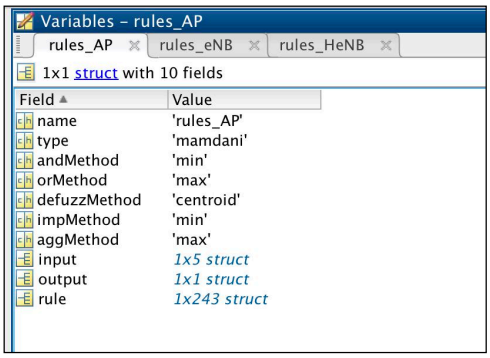


Figure 13: WiFi AP's FLC details

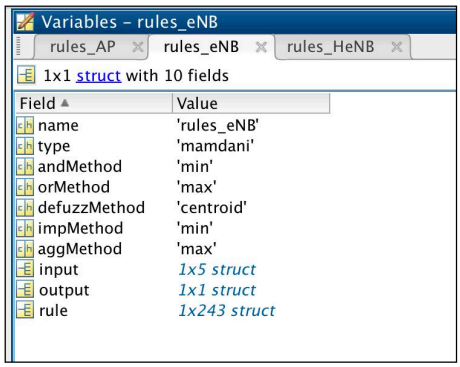


Figure 14: eNB's FLC details

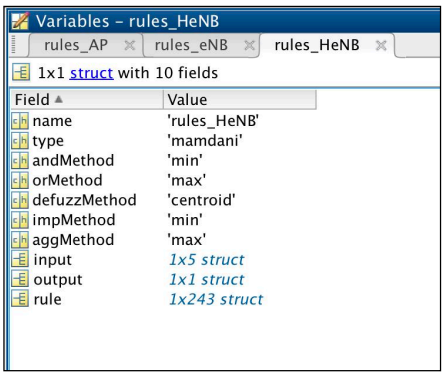


Figure 15: HeNB's FLC details

Fuzzy Inference Processes comprise five primary parts:

- Fuzzification of the input variables
- Application of the fuzzy operator (AND or OR) in the antecedent
- Implication from the antecedent to the consequent
- Aggregation of the consequents across the rules
- Defuzzification

According to the above figures (Figure 13-Figure 15), the primary options, which are set are:

- *Mamdani*-type[5] inference method is deployed. Mamdani-type inference, as defined for the toolbox, expects the output membership functions to be fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable that needs defuzzification.
- Aggregation method is max
- Implication method is min
- Centroid defuzzification method (center of gravity for the resulted final shape, after all all rule-based initial shapes are superimposed)

Each of the inputs is being fuzzified to low, medium, and high membership functions (MFs). All the inputs apart of the service sensitivity to latency are being fuzzified using triangular and trapezoidal MFs for exploiting the fact that each state is in general well defined, and has zero values for each state. On the other hand, for the service sensitivity to latency, we have used Gaussian input MFs, for highlighting the fact that all the states

[5] Yager, R. and D. Filev, "Generation of Fuzzy Rules by Mountain Clustering," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, pp. 209-219, 1994.

have non-zero values; in other words, even for the services that are considered as non-latency sensitive the user still has (loose) latency requirements.

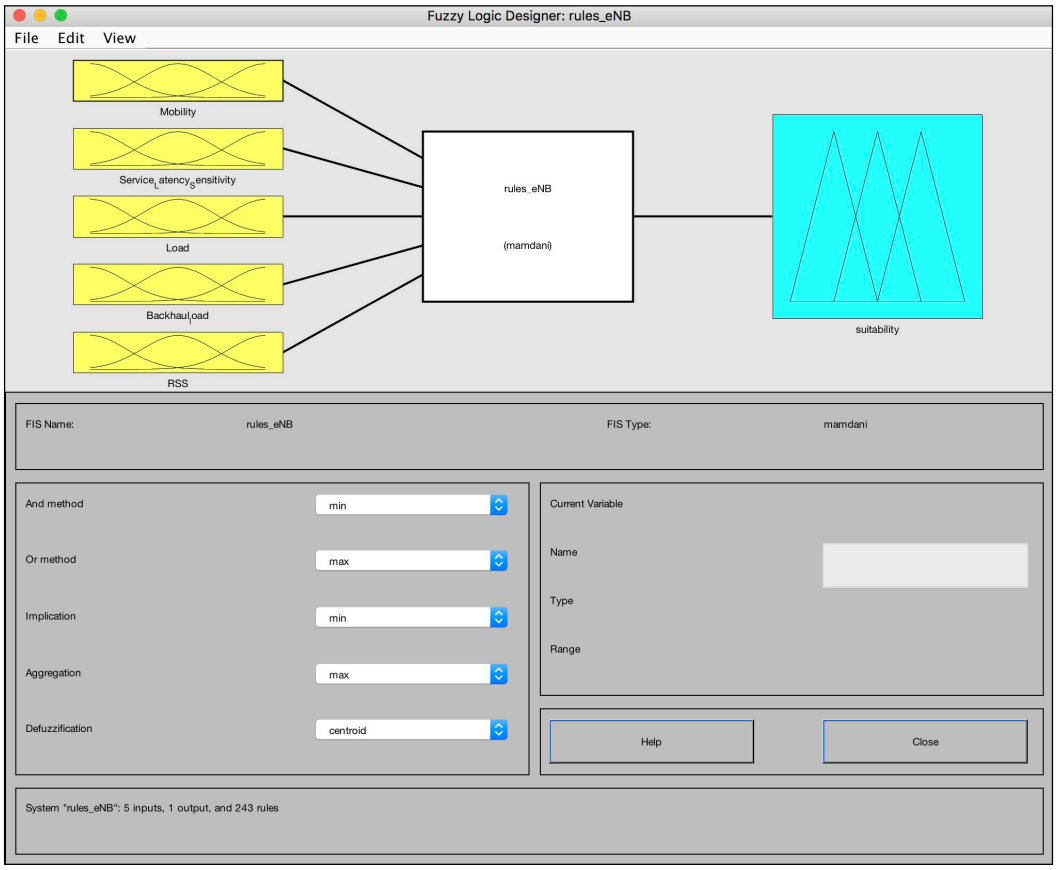


Figure 16: Fuzzy Logic Designer in MATLAB's Fuzzy Logic Toolbox

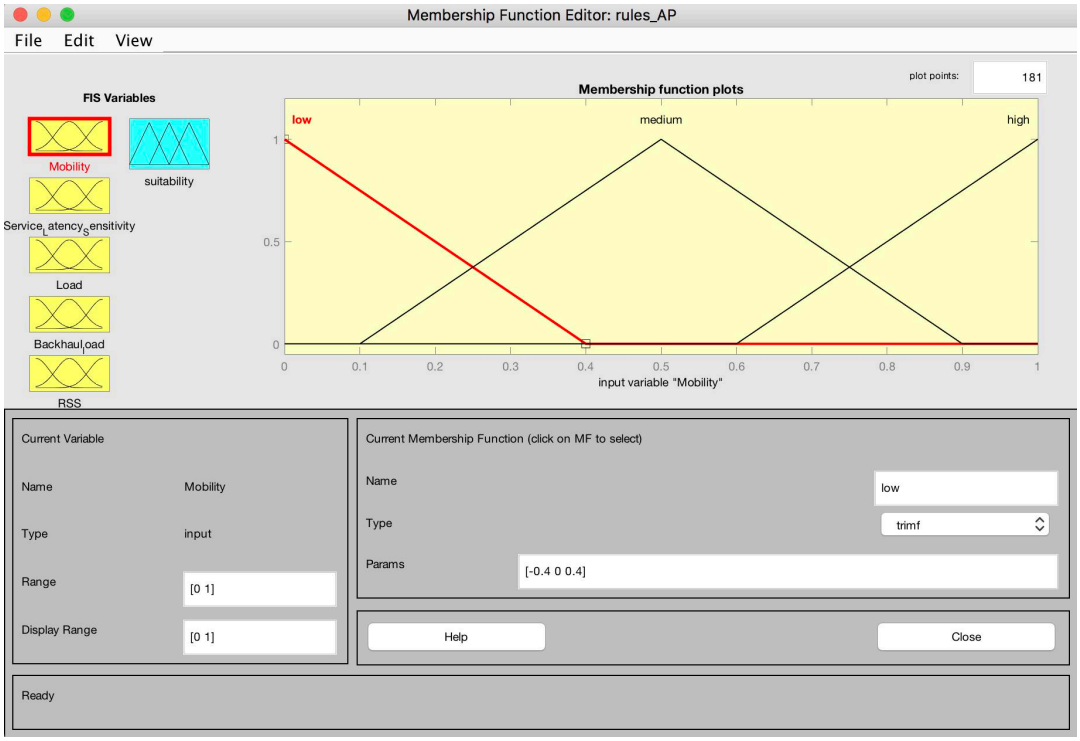


Figure 17: Example of Mobility MF for WiFi AP in MATLAB's Fuzzy Logic Toolobox

The inputs are being combined in the FIS as defined by the rule sets that have been defined in the following format:

```
IF (RSRQ == high)
AND (load == medium)
AND (backhaul_load == low)
AND (mobility == low)
AND (sensitivity_to_latency == low)
THEN Suitability = high
```

In this case for every of the 3 fuzzy reasoning modules we have defined 243 rules to cover all the potential input combinations, resulting in 729 rules overall. As also discussed earlier, these rules are subject to adaptations, according to the administrator's policies.

In order to elaborate on the input parameters and the fuzzification rules and provide a comprehensive picture of the way they influence our system, we illustrate how the input context parameters are fuzzified inside the FLC, as well as how the output is defuzzified as RAT Suitability, by providing a number of indicative 3D surfaces (Figure 18-Figure 25) that have resulted from the aforementioned rule sets. The *Suitability* value ranges from 0,0 to 1,0 (0-100% respectively) in the vertical axis. It must be noted, in addition, that each plot illustrates only two out of the overall five inputs at a time -for visualization reasons-.

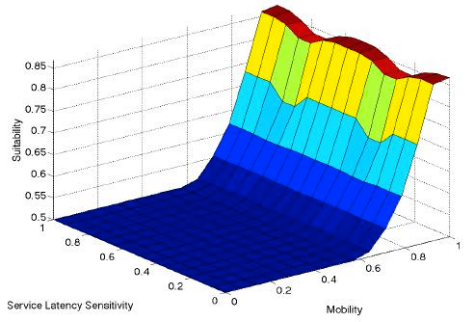


Figure 18: Suitability$_{macro}$ = f (Latency sensitivity, Mobility)

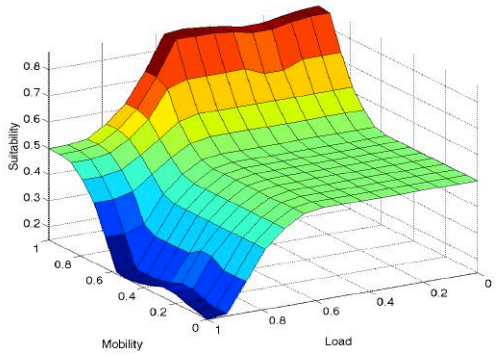


Figure 19: Suitability$_{macro}$ = f (Mobility, Load)

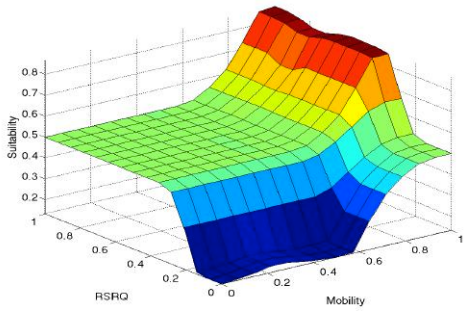


Figure 20: Suitability$_{macro}$ = f (RSRQ, Mobility)

As illustrated above (Figure 18), the *Suitability* of a macro cell ranges from 50% up to 85% (0.5 and 0.85 respectively on the z axis) for varying mobility and service latency. Specifically, it can be seen that a macro cell becomes radically more suitable when the mobility of the UE increases. This means that the proposed scheme prioritizes the placement of high velocity UEs to big cells, due to their advantage over the small cells in relation to the considerably higher time frame a fast UE may remain within a macro cell's boundaries. Lower time within a small cell's boundaries leads to increased delays that are introduced by the connection establishment procedures upon a handover. The traffic type's influence on the Suitability is almost negligible, when comparing to the mobility. Similarly with Figure 18, the UE mobility has the identical impact with the previous

example in Figure 19. Contrary to the previous case, in which the mobility played the dominant role, this time the 2nd parameter, -the traffic load of the base station-, plays a major role, causing the Suitability to decrease considerably when the load increases, and vice versa. In addition, it can be observed that when the load is less than 60% (meaning that most probably the base station is capable of serving a new UE session), only mobility influences the final outcome. Figure 20 shows how RSRQ influences the Suitability of a macro cell as well. Once more, the mobility's impact is explained in accordance with the two previous surfaces. One detail that is worth discussing is that all RSRQ values above 20% (of the max RSRQ) are acceptable; however, below that threshold the cell becomes inappropriate due to very poor connection quality, and the Suitability falls instantly for any type of UE mobility.

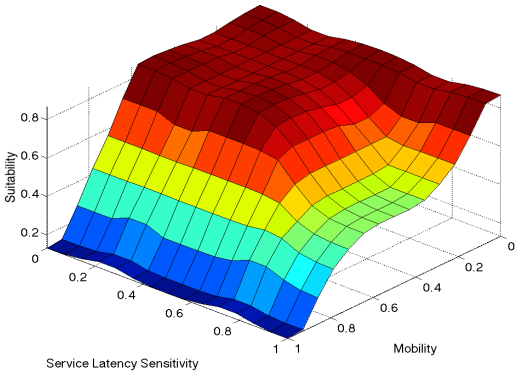


Figure 21: Suitability$_{femto}$ = f (Latency sensitivity, Mobility)

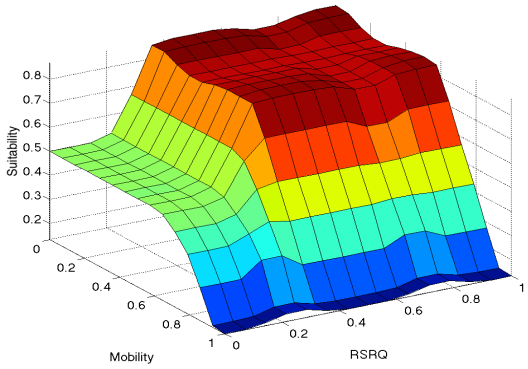


Figure 22: Suitability$_{femto}$ = f (Mobility, RSRQ)

Contrary to the macro cell's selection criteria, the higher the mobility of the UE is, the lower a femto cell's *Suitability* becomes. The reason behind this radical decrease –as it is illustrated in Figure 21- is due to the limited time frame a UE may remain within a small cell's boundaries, leading to increased delays that are introduced by the connection establishment procedures upon consecutive handovers. In other words, the proposed scheme mostly selects to offload traffic to femto cells by priority to static or slowly moving UEs. Figure 22 shows the same pattern with the respective macro cell's surface (Figure 20). The main difference relates to the inversed mobility's impact, as also discussed earlier.

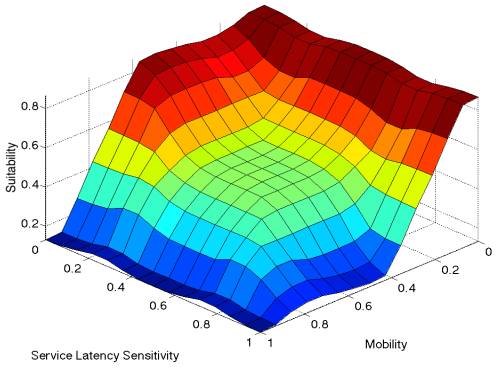


Figure 23: Suitability$_{WiFi}$ = f (Latency sensitivity, Mobility)

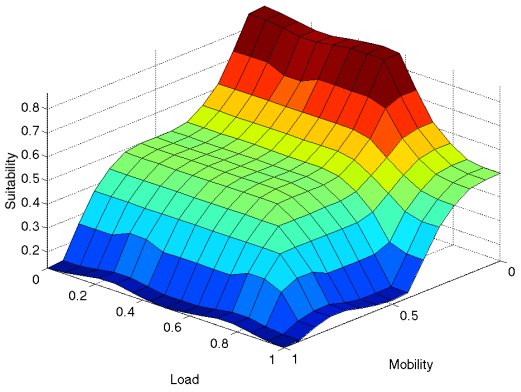


Figure 24: Suitability$_{WiFi}$ = f (Load, Mobility)

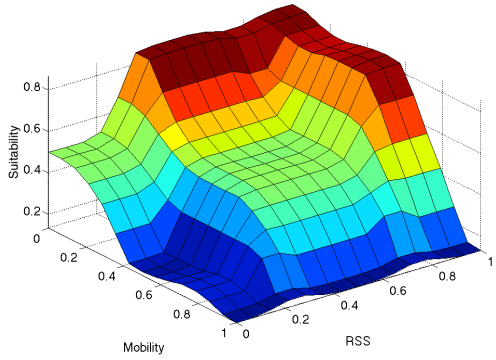


Figure 25: Suitability$_{WiFi}$ = f (Mobility, RSS)

The selection criteria of WiFi APs resemble the femto cells' due to their limited range and –as a result- the time a high mobility UE may remain within the cell's boundaries and avoid consecutive handovers or ping-pong effects. In addition, WiFi APs are suitable for serving services with high sensitivity to latency (Figure 23) due to the number of resources usually shared among fewer users, resulting in more effective resource scheduling among them. In accordance with the aforediscussed parameters and their impact on the respective RAT Suitability, Figure 24 and Figure 25 illustrate the Load – Mobility and Mobility – RSS impact on the WiFi AP's selection *Suitability*.

The defuzzification process comprises aggregating the outcomes of all the applying rules and ending up to a certain degree of the final output value, i.e., RAT *Suitability*. The defuzzification process uses MFs for capturing the degree that the output belongs to a specific state. In this case, Gaussian MFs have been used, for exploiting the smooth (i.e. the *Suitability* should be related to the inputs in a smooth manner without non-linear alterations) and non-zero (the decision maker needs to conclude to a decision based on all inputs' range) nature at all points. The RAT/cell with the highest FL output (i.e., *Suitability* value) is being selected for starting the procedure of the handover; in case of a rejection the algorithm traverses the list downwards (2nd, 3rd option, etc.). Based on the above surfaces, the overview table that follows provides a comprehensive description of the relationship between the monitored context parameters and the output *Suitability* metric (Table 9, 1st column). In the second column, we depict the correlation between different input parameters, i.e. how the impact of a specific parameter on the *Suitability* may be influenced by the value of one of the rest of the parameters.

Table 9: Overview of the system's input parameters and their impact on the system

Monitored context parameter	***Impact on Suitability***	***Correlation with other input parameters***
BS/AP load	Inversely dependent (higher load, lower *Suitability*)	a) Directly linked to the backhaul load b) The higher the RSS/RSRQ the lower the impact of load on the *Suitability*
Backhaul load	Inversely dependent	a) Directly linked to the BS/AP load b) The higher the RSS/RSRQ the lower the impact of backhaul load on the *Suitability*
UE mobility	Inversely dependent for small cells (femto/Wi-Fi) Proportional for large cells (higher mobility, higher *Suitability*)	Linked with the traffic flow type: the more sensitive to latency, the higher the impact of the mobility on the *Suitability*
Traffic flow type/ Sensitivity to latency	Inversely dependent for small cells (femto/Wi-Fi)	a) Linked with the UE mobility (see above)

		b) Influenced by the RSRQ/RSS: the higher the RSS, the less *Suitability* is influenced by the Sensitivity to Latency
RSRQ/RSS	Proportional (higher RSRQ, higher *Suitability*)	N/A

Summarizing, the main rationale behind the rules composition is the following. The UE tends to consider as optimal points of attachment the RATs/layers that are being sensed with high RSS/RSRQ, if they are not heavy loaded, or their backhaul links are not loaded. On the other hand, regardless if a point of attachment is being sensed by the UE with high RSS/RSRQ, if it is loaded (both access or backhaul link), it is unattractive for high data rate services. The UE mobility parameter influences the decision on whether a macro or a smaller (pico, femto, WiFi AP) cell will be selected. Pico/femto cells are avoided by high mobility UEs reducing this way the consecutive and redundant handovers. On the contrary, static UEs preferably are linked to smaller cells, in order to offload the traffic of the macro ones, so that they are available for the high velocity users. Finally, the higher the sensitivity to latency of the specific traffic flow, the more the mobility of the UE influences the *Suitability.*

4.6 Signaling – related issues

There is a compelling trade-off concerning all context-based schemes: on the one hand, the more information that is aggregated and processed, the higher the potential to build a comprehensive and holistic context for the user; at the same time, the drawbacks of both the burden that is placed on the network for exchanging this excessive information, as well as the computational costs on the processing entity (core network, UE, etc.) should in no way be overseen. It may be argued that if additional context inputs are taken into account, higher granularity context about each RAT/cell may be generated: Latency, coverage, RTT, number of retransmissions, BER, SINR, packet loss, throughput, bandwidth, network jitter, user monetary budget, location are all additional context parameters, which could possibly add information in the decision-making process. However, it must be very carefully taken into account that excessive context information acquisition is tightly associated with excessive signaling, posing as a consequence a redundant burden on the network, particularly when referring to very dense environments comprising thousands of UEs and tens/hundreds of available RATs. Evidently, in our scheme, the required input information for extracting the needed context has been selected carefully taking into consideration the crucial target of minimal signaling overhead; apart from the enriched ANDSF management we assume that includes the traffic load of the base stations and their respective backhaul link's, the rest of the input information required by COmpAsS is already available by existing 3GPP-stanadradized messages and interfaces. As a result, the signaling overhead imposed by the proposed scheme comprises potentially a) the signaling required for acquiring all the input parameters at the UE side (O_{input}) and b) the *Suitability* output reporting (O_{output}) back to the core network's final decision-making entity.

We assume that O_{output} equals zero, as the *Suitability* output report of the UE plays an identical role to the current 3GPP's *UE Measurements Report* consisting of the reported RSRQ values. As a result, input signaling cost is directly linked to the five input parameters (**Equation 1**)

$$O_{overall} = O_{input} + O_{output} = O_{input}$$

$$O_{overall} = \sum_{m=1}^{U} (\sum_{n=1}^{R} (S_{RSRQ} + S_{mobility} + S_{flowtype} + S_{load} + S_{backhaul}))$$

Equation 1: Signaling Cost

U is the overall number of UEs in a particular network environment. *R* is the overall number of the available RATs, which are evaluated by the UE. S_X refers to the additional signaling payload of the respective input parameter. According to the earlier analysis, 3 out of the 5 input parameters ($S_{RSRQ/RSS}$, $S_{mobility}$ and $S_{flowtype}$) actually correspond to zero overhead as they are already available in the UE according to the current 3GPP standards. The traffic load information both for the access, as well as the backhaul link are described in a coarse manner (low, medium, high); as a result, we append the additional required number of bits in the signaling messages. The traffic load information both for the access, as well as the backhaul link are described in a coarse manner (low, medium, high); as a result, we append the additional required number of bits in the signaling messages.

The figure that follows (Figure 26) illustrates how the signaling increases for the legacy RSRQ/RSS-based mechanism and the proposed one in several 5G use cases. The evaluation has been made for one evaluation of all available RATs by all the UEs in an area. To further elaborate, we consider some of the most challenging 5G use cases, which follow the METIS project specifications [182]. According to these specs, each use case corresponds to a different number of a) UEs and b) RAT choices, in the specific network environment; virtual reality office, shopping mall, traffic jam, and stadium/open-air festival are some key examples, which demonstrate an increasing number of users and available RAT options in a specific area, ranging from a few UEs and RATs in an a area (Virtual Reality Office) to thousands (Stadium).

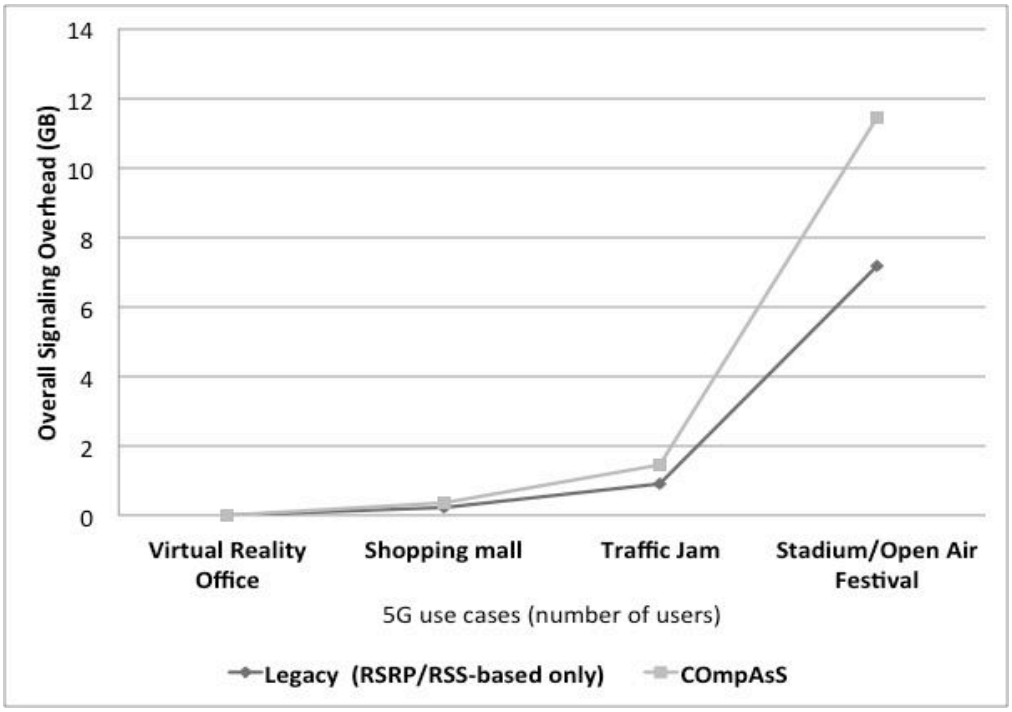


Figure 26: Signaling overhead for advanced context acquisition in diverse 5G use cases

Evidently, there is an increase in the signaling burden for acquiring the additional context parameters from the ANDSF in the case of COmpAsS. This is the case for any context-based mechanism that attempts to walk one step further from the simple RSRQ-RSRP based solutions. According to our knowledge, there is no previous work, which attempts to quantify the signaling overhead of the proposed mechanism by linking it to the current 3GPP specifications. In addition, we emphasize the fact that despite the slightly increased signaling cost, the network-related KPIs that will be presented in the Experimental Evaluation (following section) demonstrate a clear superiority of the proposed scheme both in terms of delay and throughput against the RSRQ/RSS based legacy scheme, evidently compensating thus, for the extra signaling cost.

4.7 COmpAsS Experimental Evaluation

The performance of COmpAsS is demonstrated via a series of simulation scenarios. In this section, 3 rounds of experiments will be presented, along with the respective assumptions, topologies and outcomes.

All simulations were carried out using the open-source NS-3 discrete-event network simulator ([183]). The implementation of COmpAsS inside NS-3 resulted in a custom build of NS-3.19 release, including *fuzzylite,* a fuzzy logic control library written in C++ language.

In all scenarios, which will follow, COmpAsS is juxtaposed against A2A4 RSRQ mechanism (Figure 27) –a well-established handover algorithm found often in the literature-. A2-A4-RSRQ may be triggered by the two events; Event A2 is defined as the situation during the serving cell's RSRQ becomes worse than a *threshold*. A4 event describes the situation when a neighbor cell's RSRQ becomes better than a *threshold*.

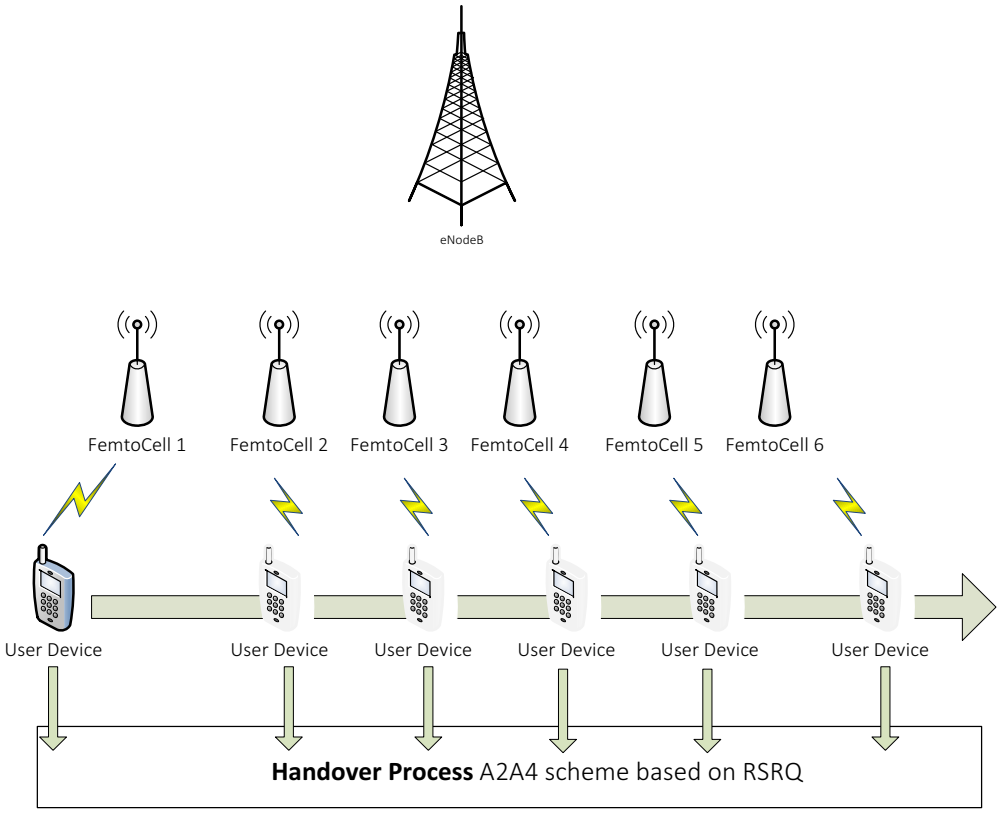


Figure 27: A2A4 RSRQ based Handover mechanism

Furthermore, throughout the experiments specific Key Performance Indicators (KPIs) were chosen, which reflect the performance of RAT selection scheme from diverse

perspectives and provide a holistic picture of the results. The pool of KPIs, which are used comprise a) the number of overall handovers which took place during the simulation, b) the downlink and uplink system throughput, c) the uplink and downlink delay, as well as the downlink and uplink packet loss throughout the experiment duration.

Last but not least, in all scenario the Shopping Mall 5G use case has been selected ([184]). We selected this set-up, as a typical setting for a future extended rich communication environment, involving both “traditional” radio networks, as well as wireless sensor networks, where customers access mobile broadband communication services while they are directly addressed by personalized location-based services of the shopping environment.

4.7.1 Experiment 1: Simplified Shopping Mall use case

The first scenario presents a simplified version of a shopping mall test case; two rows of femto cells (assuming they are inside the mall’s shops) and a pedestrian corridor in the middle, while macro cells (LTE eNBs) co-exist at a distance of around 1km which is a typical range for an urban – suburban location (Figure 28). In our simulation scenarios, we included 2 eNBs and 5 HeNBs. For simplicity’s sake, no WiFi APs were used in the scenario, as the large-small cell handling evaluation of our mechanism was achieved by using LTE macro and femto cells. It is furthermore assumed that inside the mall area, several UEs are either static or moving at pedestrian’s speeds, i.e. 0 – 1.5 m/s. These UEs –being attached to the mall’s HeNBs- contribute as well in the creation of the load that needs to be taken into account for selecting the appropriate RAT from UE.

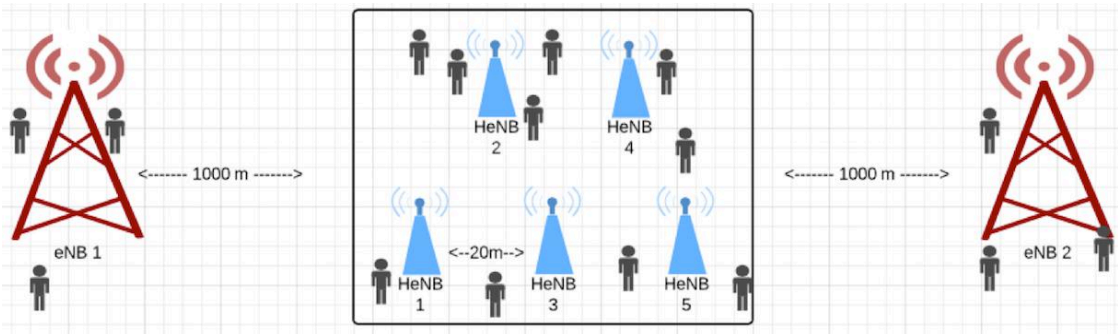


Figure 28: Simple topology (Experiment 1) towards proof of concept

The simulations were made using three moving UEs at pedestrian speed: low, medium and high. In order to evaluate the performance of the algorithms, we increased the load of two out of the five overall HeNBs gradually, reaching from zero load to very high load. The background traffic in all rest (H)eNBs causes them to be in a medium load state during simulation time. By load, we are referring to both Load and Backhaul Load –as they were presented in the previous sub-sections-, which is calculated by the number of the rest static UEs, associated to each (H)eNB, as well as the number of bearers per UE. The table below summarizes the simulation parameters:

Table 10: Experiment 1 simulation details

eNB (macrocell)		**HeNB (femtocell)**	
Number of eNBs	2	Number of HeNBs	5
Carrier frequency (MHz)	Downlink: 2120.0 Uplink: 1930	Carrier frequency	Downlink: 2120.0 Uplink: 1930
Channel bandwidth	50 RBs	Channel bandwidth	15 RBs

Transmit power	35.0 dBm	Transmit power	23.0 dBm
Rest parameters			
Simulation time		225 s	
Simulation time unit		0.1 s	
Test UEs Sensitivity to Latency		High (0.7/1.0)	
UE mobility		UE 1: linear low velocity (0.4 m/s) UE 2: linear medium velocity (0.8 m/s) UE 3: linear high velocity (1.4 m/s) Rest attached UEs: static close to the associated (H)eNBs	

The KPIs, which are assessed, are the overall number of handovers that took place, the average throughput as well as the average experienced delay for the three moving UEs.

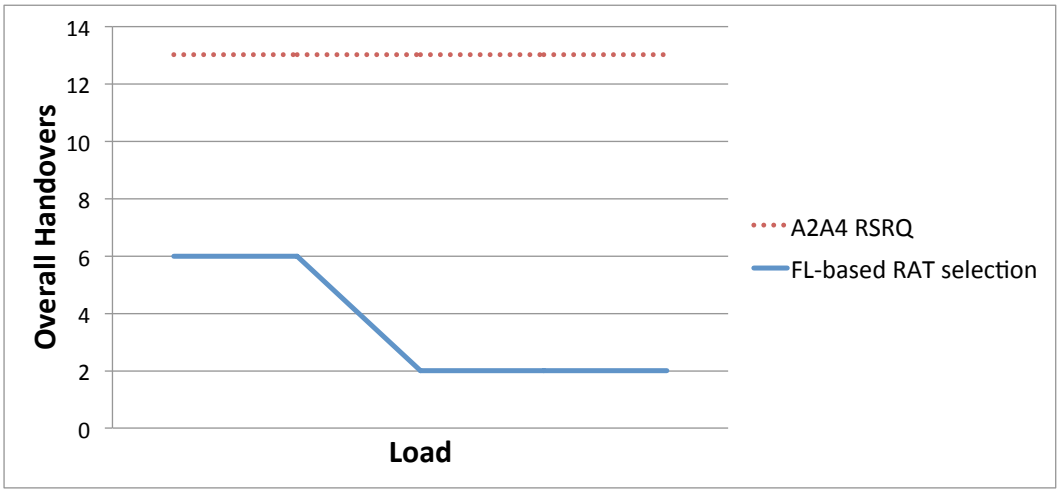


Figure 29: Overall number of handovers that took place

Figure 29 illustrates the performance of the two algorithms in terms of the overall number of handovers that took place from the 3 UEs. Noticeably, the A2A4 RSRQ algorithm's decisions are not influenced neither by the higher mobility of the UEs, nor by the increasing load of the HeNBs. On the contrary, the proposed mechanism tends to minimize handovers in the afore-mentioned cases. When the load is low, the number of handovers is reduced by 53.8% -due to reduced number of executed handovers of the high mobile UEs-, while when the load increases, the overall handovers are reduced by 84.6% since the suitability factor of candidate (H)eNB is low.

Figure 30 illustrates the performance of the algorithms with regard to the calculated average throughput both in the downlink and the uplink. The FL-based RAT selection outperforms the A2A4 RSRQ in terms of throughput; both in the downlink, as well as the uplink case by an offset of 300 kbps roughly as load increases. An interesting observation in the case of the proposed mechanism's performance is the increase of the throughput when the load of the HeNBs is extremely high; this advantage results from the fact that as load increases the suitability of the femto cells is constantly decreasing, tending to retain the UE from doing a handover to them. As a result, throughput and delay performance are directly related to the handover decisions presented earlier.

Similarly, the difference in the delay (measured as A2A4 RSRQ average packet delay minus the FL-based average packet delay) remains positive in all scenarios. Once more,

as load increases radically, the FL-based mechanism tends to increase its performance since the UE keeps its connection to medium loaded eNB.

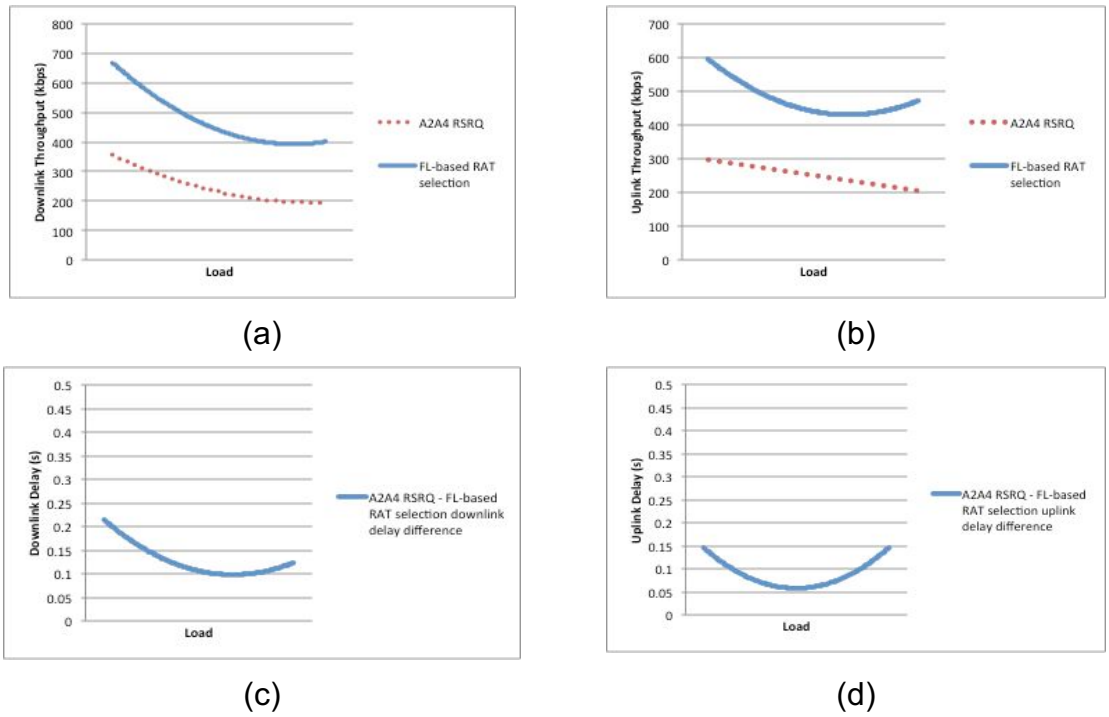


Figure 30: (a) Downlink Throughput, (b) Uplink Throughput, (c) Downlink Delay, (d) Uplink Delay

4.7.2 Experiment 2: Realistic Shopping Mall use case

The 2nd experiment presents a realistic business case scenario of a shopping mall comprising 3 floors (ground floor, 1st and 2nd floor), and 20 shops per floor (Figure 31). The UEs are either static or moving, and are roaming around the shopping mall rooms (shops, cafes, etc.). Several HeNBs are deployed in the three floors. In addition, two macro cells (eNBs) exist outside the mall area in a distance of 200m to different directions. Due to the fact that COmpAsS handles Wi-Fi APs and HeNBs in a similar way, with regard to the pre-defined rules of the Fuzzy Inference Engine, for the sake of simplicity, in the simulations only macro and femto-cells are deployed.

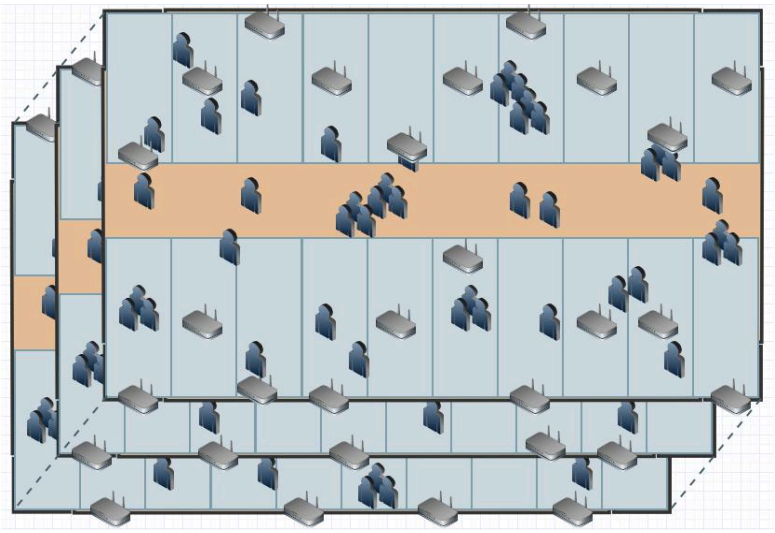


Figure 31: Experiment 2: Shopping mall with 3 floors and 20 shops per floor

Besides the several UEs, which are roaming inside the mall area and creating respective traffic to the HeNBs, we use one “test UE”, in which COmpAsS is deployed. Different

simulations were carried out to test the UE at different velocities (low, medium, high), in each one of the scenarios in order to evaluate the proposed scheme for varying UE mobility, as mobility is one of the inputs, which are taken into consideration for the decision. The test UE is moving with linear velocity between the rows of the shops, on the 1st floor. An overview of the simulation details is presented in the following table:

Table 11: Experiment 2 simulation details

Environment	Shopping mall: 3 floors, 100 x 200 meters per floor, 20 rooms per floor (2 rows of 10 equal rooms)
Number of UEs	Variable (UEs connecting/disconnecting)
Number of (H)eNBs	2 eNBs, 9 HeNBs
Carrier frequency (MHz)	Downlink: 2120.0, Uplink: 1930
Channel bandwidth	50 RBs for eNBs, 15 RBs for HeNBs
Transmit power	35.0 dBm (eNBs) , 23.0 dBm (HeNBs)
Simulation time	100 s
Time unit	0.1 s
UE mobility	0.4 m/s, 0.8 ms, 1.4 m/s (linear constant velocity)
HeNB load	Varying depending on the number of associated UEs (very low, low, medium, high, very high)
Traffic sensitivity to latency	High (0.7/1.0)

The following figures illustrate the measured KPIs, which resulted from the two mechanisms with regard to the number of overall handovers which took place during the simulation, the throughput of the test UE, the experienced delays, as well as the packet loss during the measurements.
Variable load of the femto-cells of the shopping mall was tested, calculated in relation to the overall associated users per base station and traffic that is generated. In particular, the load of the base stations varies from 10% up to 90% of their available resources (horizontal axis in Figure 32- Figure 38).

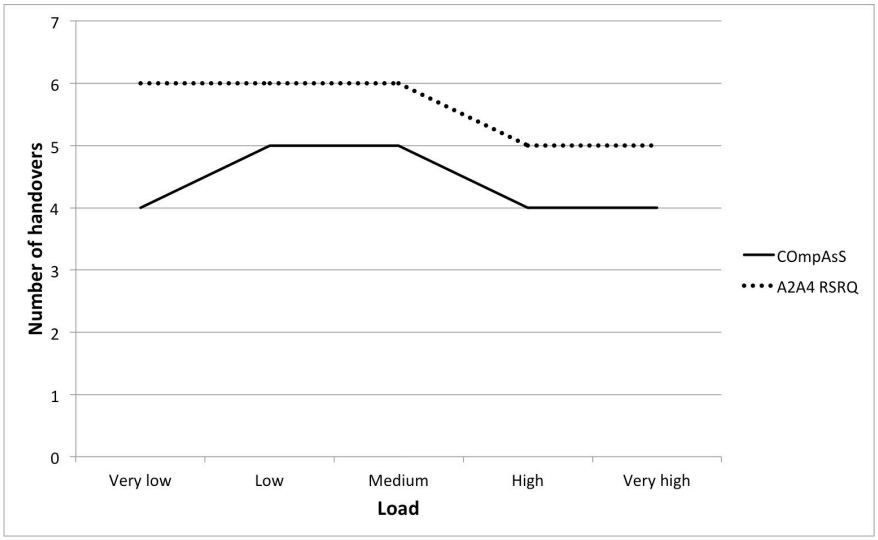


Figure 32: Number of handovers

In Figure 32, the overall number of handovers is shown. According to the graph, the proposed mechanism tends to minimize the number of handovers as it realizes less handovers than A2A4 RSRQ in all load situations.

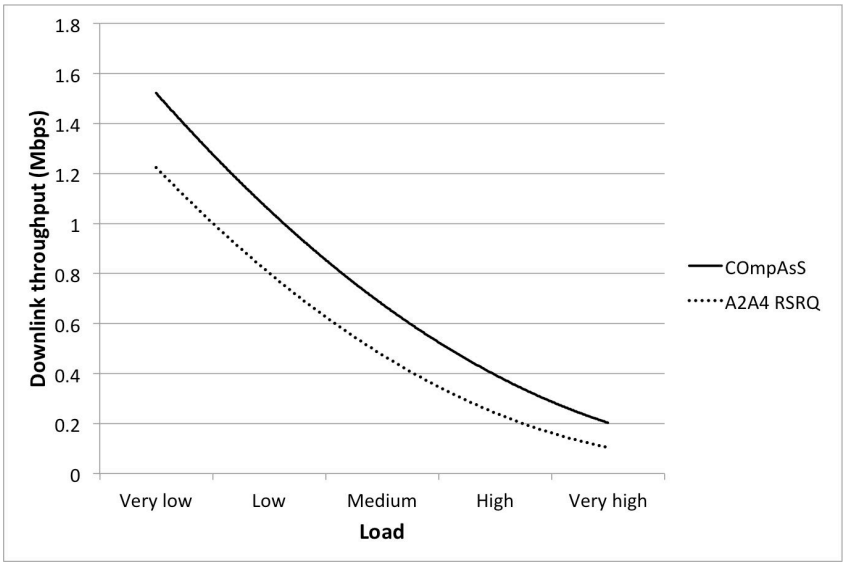


Figure 33: Downlink throughput

In Figure 33 - Figure 35, the results of the downlink are illustrated: throughput, delay and packet loss. With regard to the throughput (Figure 33), COmpAsS outperforms the A2A4 RSRQ algorithm in all load scenarios by 10-20 %. In the case of the proposed scheme, the high interference, which results from the tested environment retains the UE from handing over to the femto-cells, which suffer more; instead, the UE tends to stay more time attached to the eNBs, achieving finally a higher throughput. Moreover, the UE mobility is taken into consideration from COmpAsS, in contrast to A2A4 RSRQ; for high mobile users femto-cells are less attractive, particularly if the load of them increases as well, which makes them even more unattractive. In the case of the delay (Figure 34), a significant difference between the two mechanisms is observed throughout the measurements. Similarly, the packet loss (Figure 35) that experiences the UE, which uses the COmpAsS mechanism, is by 20% lower than the other scheme, no matter how high the load of the network –and as a result the experienced interference as well- is.

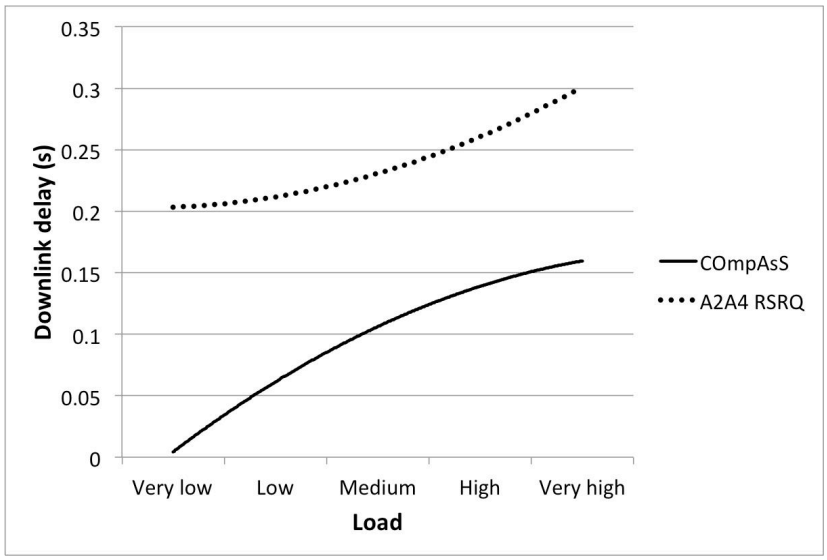


Figure 34: Downlink delay

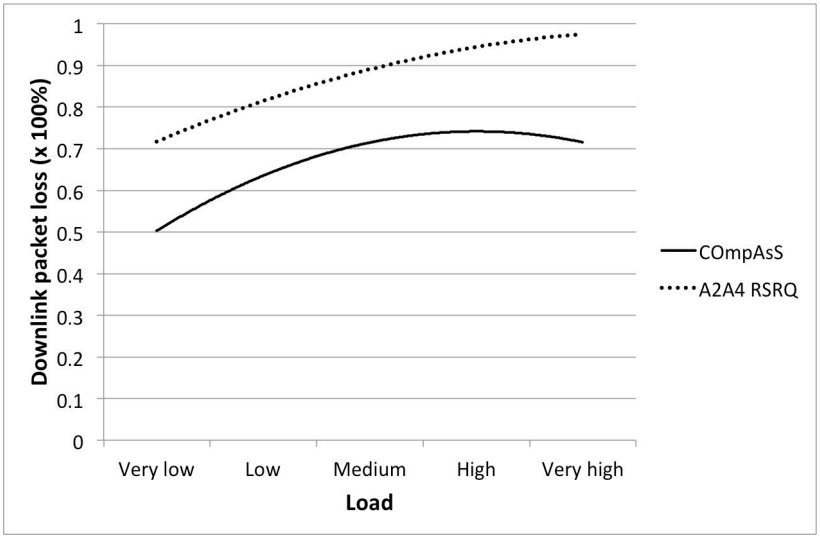


Figure 35: Downlink packet-loss

Figure 36 - Figure 38 illustrate the measured KPIs of the uplink. Noticeably, the difference of the throughputs of the two schemes is even higher than in the case of the downlink, i.e., 200 – 400 Kbps (Figure 36).

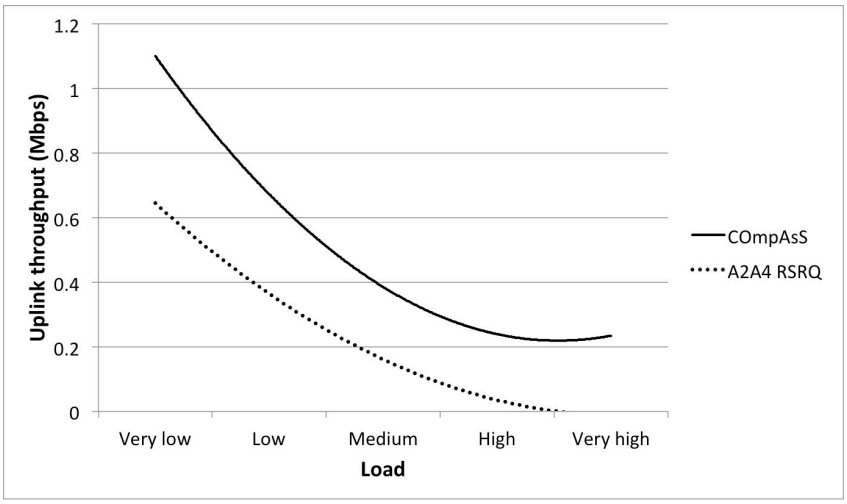


Figure 36: Uplink throughput

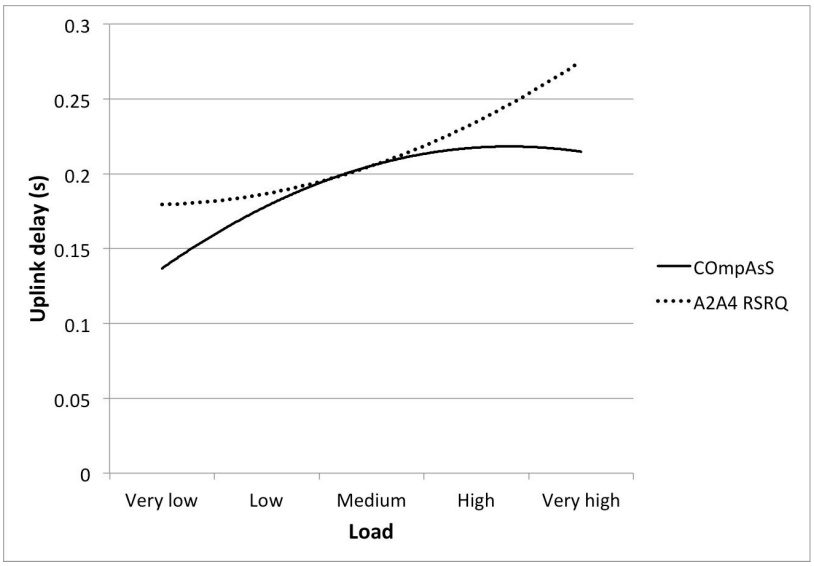


Figure 37: Uplink delay

With regard to the uplink delay (Figure 37), it is shown that, although at medium load the two algorithms have almost identical results, as the load increases further, COmpAsS's

performance is significantly better –roughly 50ms-, maintaining constant delay. In contrast, A2A4 RSRQ's delay is increasing further. This is explained by the fact that, the suitability by COmpAsS during the load increase of the femto RATs, reduces radically, particularly for faster users.

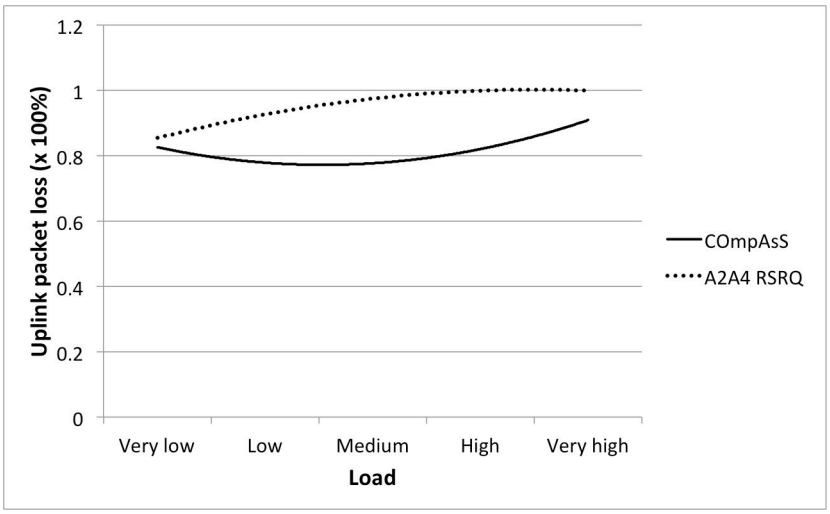


Figure 38: Uplink packet-loss

The packet-loss in the uplink case (Figure 38), similarly with the previous figures confirms the superior performance of the proposed mechanism.

4.7.3 Experiment 3: Advanced Shopping Mall use case

The final COmpAsS evaluation experiment is the most advanced one, among the three, which are included in this evaluation section.
We evaluate this 3rd experiment on the basis of 4 different scenarios, which we describe in detail below. Overall, the network deployment allows seamless handling of services across different domains, e.g. mobile/fixed network operators, real estate/shop owners and application providers. The environment is similar to Experiment 2, presented earlier (Figure 39). Each floor's dimensions are 200x100m, containing 20 rooms/shops per floor, with several LTE Femto cell placed on each floors, depending on the scenario. Outside, two LTE eNBs are placed, 150m north and west of the mall respectively.

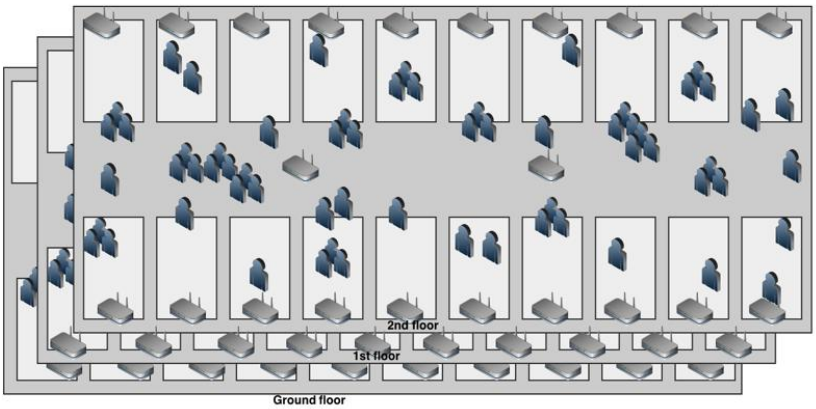


Figure 39: Experiment 3 network environment

In order to evaluate the proposed framework, using LTE femto cells and macro cells is sufficient, as the rules that apply for Wi-Fi are almost identical with the ones applied for femto cells. In addition, the IP flow mobility between LTE and Wi-Fi networks is not available in the NS3 simulator that was selected. Our simulation scenarios are based on 3GPP Specifications [185] and [186]. In higher detail, the transmission mode is SISO (Single Input Single Output) and the scheduler is the NS-3 implementation of the

Proportional Fair MAC scheduler. We use the Hybrid Buildings Propagation Loss Model for path loss implemented in NS3 with Internal Wall Loss at 10.0 db Shadow, Sigma Indoor at 10.0 db. The network node configuration appears in Table 12. Services are implemented using NS3's UDP client-server application model and the desired data rates are achieved through configuration of the packet size and the inter-packet interval parameters. The service schedule for every user is pseudo-randomly generated at the beginning.

Table 12: Experiment 3 simulation details

NS3 Network Node	**Tx Power (dBm) [185]**	**Downlink (DL) Earfcn (MHz) [185]**	**Bandwidth (RBs) [185][186]**	**Antenna Type [185]**
Macro cell	35	2120	50 (10 MHz)	Parabolic, 15 dBi
Femto cell	20	2120	15 (3 MHz)	Isotropic
UE	20	-	-	Isotropic
Other parameters				
Number of eNBs	2			
Number of HeNBs	50 (max.)			
Number of UEs	50			
Simulation time	225 s			
Time unit	0.1 s			
Transmission mode	SISO (Single Input – Single Output)			

As already presented earlier, the proposed framework's algorithm uses two parameters, i.e. *Suitability Threshold and Hysteresis*. Different parameter values may alter radically COmpAsS's responsiveness and functionality, primarily in terms of triggering events frequency. Different network "states" (e.g., denser or scarcer deployments) would require different configurations of these two control parameters. Towards this fine-tuning process, hence, in the first two scenarios, we incorporate in our experimentation a range of values, both for Threshold and Hysteresis. Overall, the evaluation of COmpAsS moves along 4 axes-scenarios, each one of which focuses on a different varying parameter of the experiment's setup, in order to simulate -in the most realistic extent possible- all the radio conditions and network "states" that the proposed framework may encounter. In the following table, we present in more detail the 4 different scenarios.

Table 13: Experiment 3 scenarios

Scenario	**Scenario Parameter**	**Value range**	**Number of experiments**	
			COmpAsS	**A2A4-RSRQ**
1	*Suitability threshold*	[0.99, 0.9, 0.7, 0.5, 0.1]	75 experiments (15 different executions per threshold value)	15 experiments (*Suitability threshold* does not apply)
2	*Suitability margin*	[0.3, 0.2, 0.1, 0.01, 0.001]	75 experiments	15 experiments (*Suitability threshold* does not apply)

3	*Deployment density (number of HeNBs)*	[2, 5, 10, 20, 50]	75 experiments	75 experiments
4	*Network load (number of traffic bearers/UE)*	[1, 2, 3, 5, 10]	75 experiments	75 experiments

For each one of the 4 scenarios, we ran 75 similar experiments in order to maximize the validity of our experimentation results. In order to define the number of runs per sub-scenario, as well as the experiment duration, we initially carried out some test scenarios; each one of the different runs incorporates a random generation of mobility patterns for the UEs, as well as slightly varying traffic models. We defined our confidence level at 95% in order to be able to demonstrate a satisfactory and statistically valid outcome. More specifically, we calculated the confidence interval in relation to the number of runs per scenario:

We used the well-known $Za/2 * \sigma/\sqrt{(N)}$ formula, where Za is the confidence coefficient, σ is the standard deviation and N is the sample size. The standard deviation introduced a) by the step of 0,1 second of the discrete-event NS-3 simulator and b) the deviation of the received results, finally led us to execute 15 runs per scenario of 225s (2250 samples per KPI per scenario with 0.1s step), in order to reach our target –i.e., *95% of confidence level*-. As an example, in terms of actual metric values, the delay KPI is expressed in one of our scenario results as 0.13±0.00007s.

In addition to the aforementioned 300 experiments, another 180 experiments were run using the A2A4 RSRQ algorithm, resulting overall in 480 experiments, providing, thus, an extensive set of results for being able to accurately assess the validity and viability of the proposed scheme. The Key Performance Indicators (KPIs) that are used for the evaluation comprise average uplink and downlink throughput, as well uplink and downlink delay. In order to evaluate the above KPIs, we deploy randomly among the simulation UEs 4 different service categories, corresponding to 4 different traffic models, as well (see Table 14). The simulation UEs are running the respective assigned services throughout the whole simulation duration. For each one of the service/traffic flow categories, different KPIs are applied, according to the particular flow QoS requirements.

Table 14: Service/application parameters used and respective KPIs applied

Deployed Service Types, respective traffic models and KPIs		
Type	**Traffic model**	**QoS-related KPIs assessed**
1. Conversational voice	Average call duration is 1.8'[187] Average rate is 12.65 kbps[188][189]	UL/DL delay
2. Conversational video	Average video duration is 4.12' [192] Average DL speed 443 kbps [191] Average size for 480p video is 250MB per hour in [193]	UL/DL delay UL/DL throughput
3. Real-time gaming Data (non latency-tolerant)	Average packet size is 50kB Inter-packet time interval is 50 ms [194]	UL/DL delay

4. Non-GBR services (e-mail, chat, FTP, file sharing, etc.)	Average session for file download is 9.8s for 3MBs file [190]	UL/DL throughput

We derive the results of the aforementioned KPIs (DL and UL throughput and delay) by grouping each time only the UEs that are running a KPI-related service type (e.g. average DL throughput is derived only from the UEs running either conversational video or non-GBR services, according to the table).

Scenario 1: Varying Suitability threshold

In our first evaluation scenario the *Suitability threshold* ranges between two extreme values: 0.1 and 0.99; taking into consideration that in the proposed scheme, context evaluation and *Suitability* calculation procedures are performed only when the current RAT's *Suitability* has fallen below the *threshold,* the behavior of COmpAsS varies significantly, primarily as far as the triggering frequency of the mechanism is concerned. The extreme value range has been selected on purpose, targeting to demonstrate how the algorithm responds under any possible configuration. The first set of results (Figure 40-Figure 43) depicts the throughput and delay KPIs both for the uplink and the downlink.

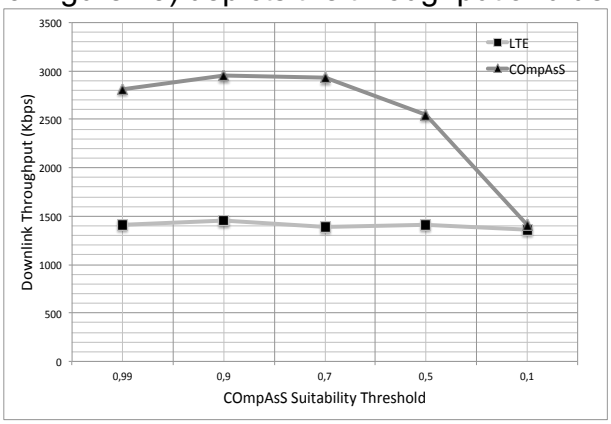


Figure 40: DL throughput for varying *Suitability Threshold*

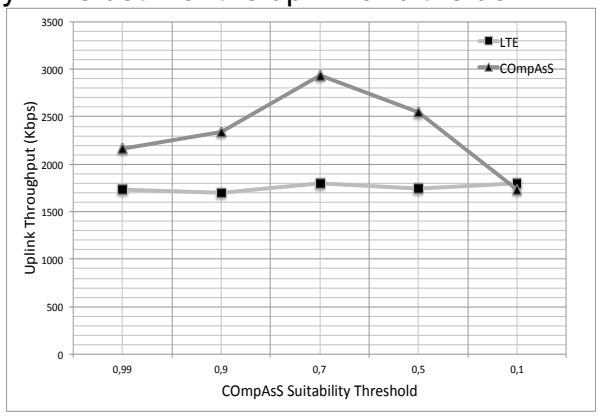


Figure 41: UL throughput for varying *Suitability Threshold*

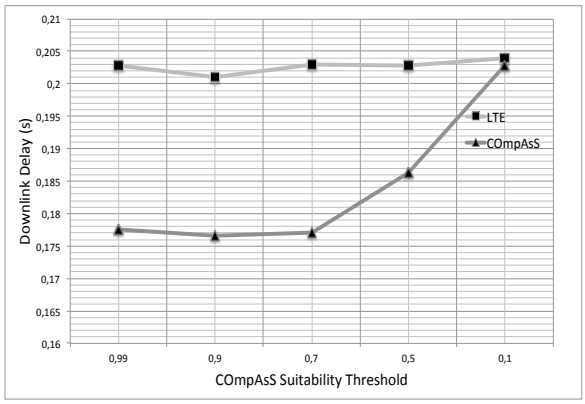


Figure 42: DL delay for varying *Suitability Threshold*

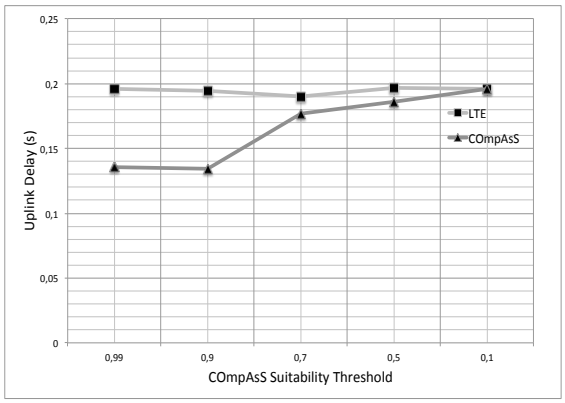


Figure 43: UL delay for varying *Suitability Threshold*

Studying the throughput graphs (Figure 40, Figure 41) reveals how COmpAsS outperforms the conventional LTE algorithm for all threshold steps, both in the downlink and the uplink respectively. In some cases, COmpAsS's performance is higher by more than 100%. Similarly, in the delay graphs (Figure 42, Figure 43), particularly for high threshold values, in the downlink, the proposed scheme's measured average delay is reduced by 10%, while in the uplink up to 25% when comparing to the A2A4 RSRQ.

High attention should be paid to the trade-off resulting from the *Suitability threshold* value, when experimenting with extreme high and extreme low values. Figure 41 demonstrates such a case, in which the optimized configuration results when *threshold* equals an intermediate value, i.e. 0,7. Extremely high values result in excessive number of algorithm triggers, which –although result in the optimal RAT selection at any point of time- cause signaling overhead and calculation latencies. In other words, for every network setting and environment, there is an optimal value for the *threshold,* for which the network metrics are the highest; this does not always correspond to the highest *threshold* value. All in all, the configuration of the system in terms of the *threshold* relies each time upon the specific use case, the administrator's targets and depends heavily on the specific environment's conditions and QoS requirements.
Nevertheless, it must be highlighted that in all graphs (throughput and delay, both DL and UL), the two mechanisms' performance converges as the *threshold* decreases, while it becomes identical for the lowest *threshold* value (i.e., 0.1). This is explained by the fact that, the lower the *Suitability threshold* value, the fewer times COmpAsS is triggered and thus, the less effective it is throughout the overall simulation duration. When the value reaches 0.1, the system responds as if the algorithm is practically no longer active.
In accordance with the above results, we select the value of 0.7 for the next scenarios of the experimentation; using the specific value, the throughput optimization is highest, while the delay remains considerably lower than the baseline, both in uplink and downlink.

Scenario 2: Varying Suitability hysteresis (margin)

The second evaluation perspective illustrates the simulation outcomes in relation to the *Suitability hysteresis,* in a similar pattern with the first scenario; the *hysteresis* ranges between two limit values: 0.3 and 0.001. As it can be inferred, the higher the *hysteresis (*i.e. the difference between the current RAT's and the candidate target RAT's *Suitability* value)*,* the "stricter" requirements of COmpAsS in terms of *Suitability* of the successor RAT for handover. On the contrary, the lowest *hysteresis* value, i.e. 0.001, implies that a handover is realized by the time a RAT with higher *Suitability* is detected, without actual margin being taken into account. The following figures (Figure 44-Figure 47) illustrate the results of the 2nd experimentation round:

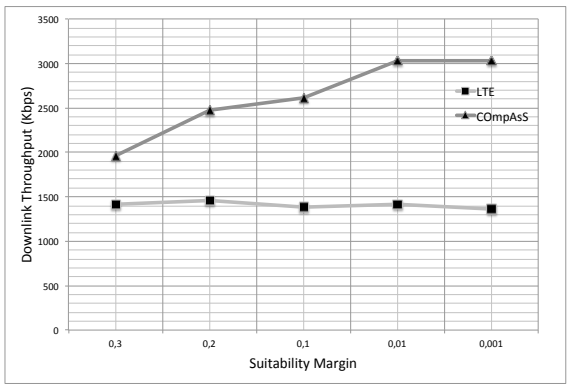


Figure 44: DL Throughput for varying *Suitability Hysteresis*

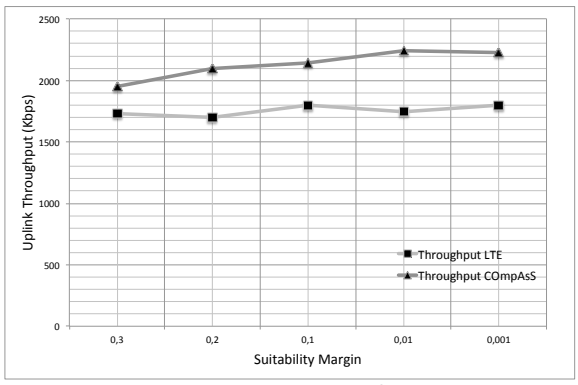


Figure 45: UL Throughput for varying *Suitability Hysteresis*

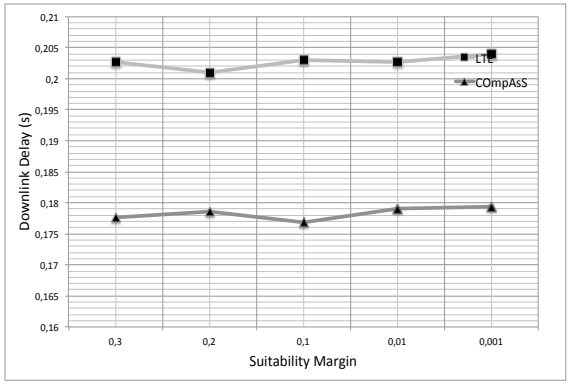


Figure 46: DL Delay for varying *Suitability Hysteresis*

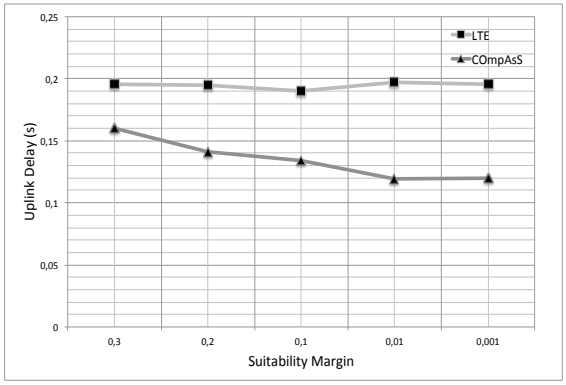


Figure 47: UL Delay for varying *Suitability Hysteresis*

It is clear that the proposed scheme's results showcase a clear superiority, both in terms of throughput and delay. The throughput is constantly increasing as the *hysteresis* is decreased, both for the uplink and the downlink. This is expected in the sense that the lower the *hysteresis,* the more handovers are realized according to COmpAsS's decision making and RAT selection. As far as the delay KPI is concerned, the proposed scheme outperforms the LTE legacy algorithm for all hysteresis values. In the downlink (Figure 46) the delay is not influenced by the hysteresis, while in the uplink (Figure 47), it follows a pattern similar to the throughput results.

It is worth highlighting the trade-off that results from the specific results. During the design of our mechanism, we introduced the *Suitability hysteresis*, aiming at facilitating the control over our scheme on the UE, in terms of battery consumption, while maintaining at the same time the highest performance possible. As it is directly inferred from the aforepresented figures, a low hysteresis implies maintaining constantly the optimal RAT selected for the UE for all flows; at the same time, this is accompanied by higher battery consumption as well (more frequent context information retrieval, FL-based Suitability metric re-calculations, etc.), as well as higher signaling overhead. Hence, depending each time on the specific use-case and respective requirements, COmpAsS can be configured in one of the two modes (i.e. low consumption or highest performance possible), as both seem to outperform conventional LTE schemes. For the rest of the experimentation scenarios, we have set the hysteresis at the value of 0.1, attempting to provide a balanced outcome when investigating the rest of the network parameters.

Scenario 3: Deployment density

In the context of the third evaluation scenario, we compare the two mechanisms in a varying environment, in terms of HeNBs' deployment density. The number of the femto cells ranges from 2 (sparse deployment) up to 50 (ultra-dense deployment); the latter resembles a typical 5G scenario as already discussed earlier. In the figures below (Figure

48-Figure 51) we illustrate the network KPIs that resulted from the third scenario's experimentation. At this point we remind the reader that the *Suitability threshold* of COmpAsS algorithm is set to 0.7, while the *hysteresis* parameter at 0.1.

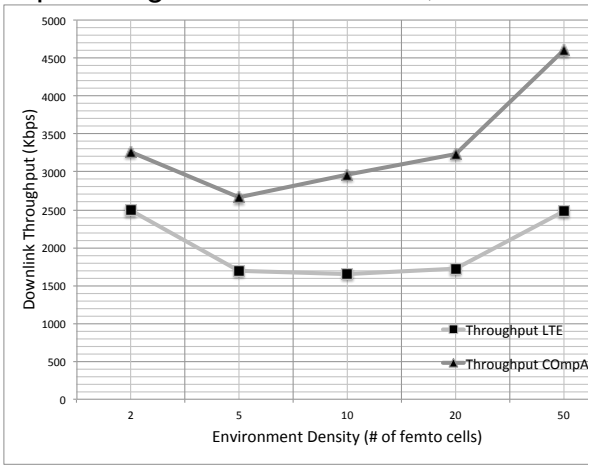


Figure 48: DL Throughput for increasing network density

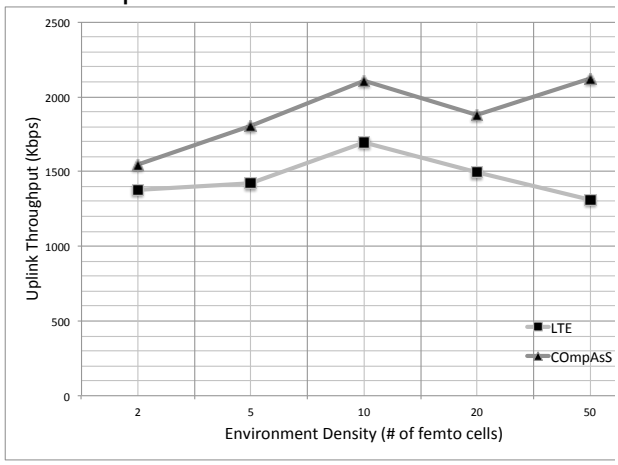


Figure 49: UL Throughput for increasing network density

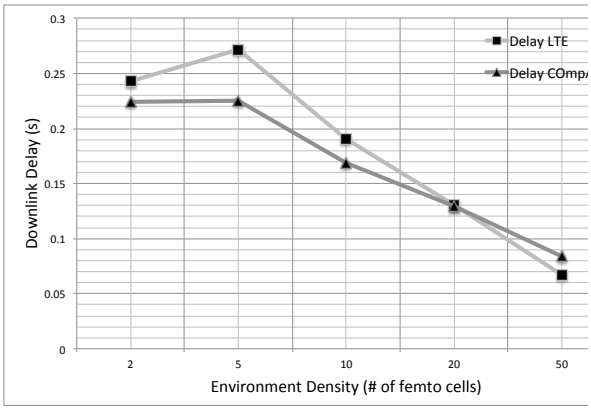


Figure 50: DL Delay for increasing network density

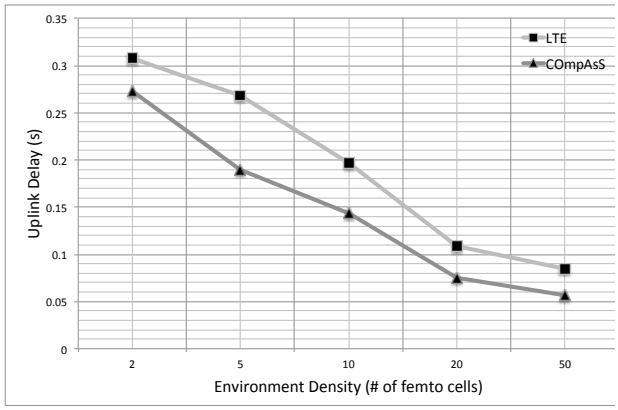


Figure 51: UL Delay for increasing network density

In terms of throughput, COmpAsS achieves higher performance both in the downlink, as well as the uplink. As also illustrated in the respective graph figures, there is no linear relation to the number of available femto cells and the achieved throughput; however, the highest throughput is achieved when the highest number of HeNBs is available. Taking into consideration, that no limitations are posed by the backhaul network, more HeNBs provide more resource blocks to the simulation UEs; hence, an efficient, context-aware mechanism such as the proposed scheme is capable of optimizing the distribution of the UEs among the femto cells to the most efficient way –in terms of resource blocks (RBs) allocation per UE- possible. Similarly, in the delay graphs, a similar performance outcome is illustrated. Both in the downlink and the uplink our scheme outperforms the A2A4 RSRQ algorithm. Once more, the high number of RBs facilitates their allocation to the respective UEs, primarily in terms of scheduling and hence, results in lower delays.

Scenario 4: Network load

In the final round of experiments and in the context of the last scenario, we gradually increase the network load in terms of active bearers (active traffic flows) per UE, aiming at comparing the performance of our scheme in extreme load and interference conditions. In the particular set-up, we deploy 10 femto cells (co-existing with the fixed –throughout all experiment scenarios- number of macro cells). Similarly with the previous scenarios, we provide the downlink and uplink throughput and delay KPIs (Figure 52-Figure 55) for the two mechanisms.

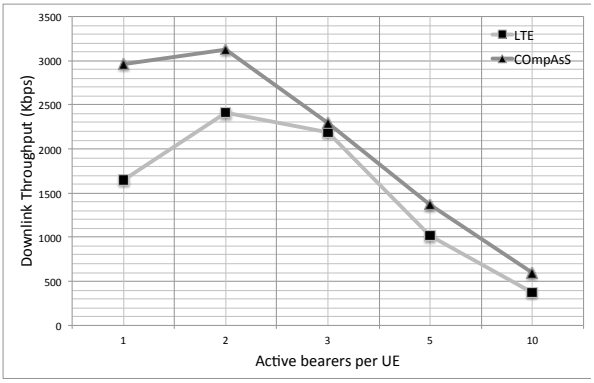


Figure 52: DL throughput for increasing network load

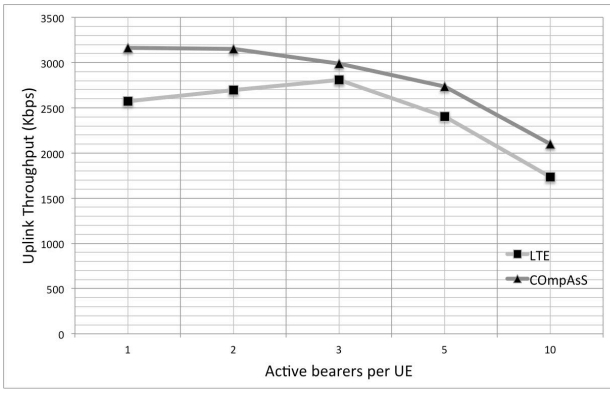


Figure 53: UL Throughput for increasing network load

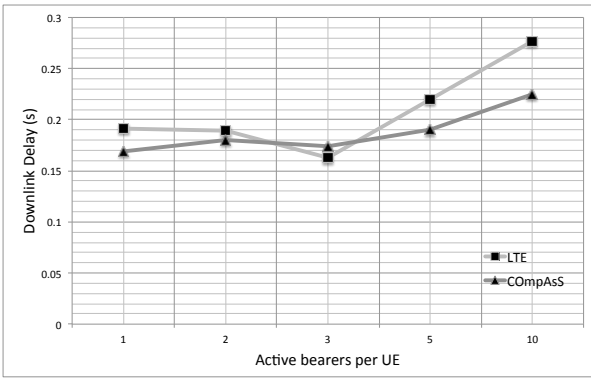


Figure 54: DL Delay for increasing network load

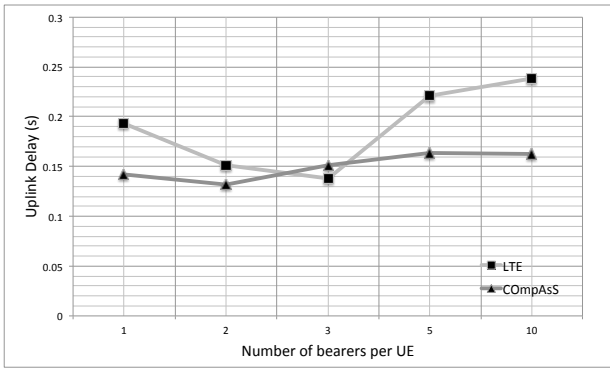


Figure 55: UL Delay for increasing network load

The performance of both mechanisms is radically deteriorated when the network load gradually increases, both in terms of throughput and delay as well, as illustrated in the above graphs. As far as the throughput metric is concerned, the proposed mechanism, -although influenced as well by the load increase-, demonstrates a superior performance than A2A4 RSRQ. This applies both for downlink and uplink. The negative impact of the load increase is illustrated primarily in the downlink throughput case, where the lack of available RBs –comparing with the overall active bearers- results in a decrease of around 80%. An analogous outcome is identified in the delay metric results, where a gradual delay increase is reported along with the load increase. Similarly with the throughput KPI, delay is also directly related with the scheduling inefficiency when the ratio of available RBs and the number of active bearers in the overall experimentation is limited.

5. CEPE: A CONTEXT EXTRACTION AND PROFILING ENGINE FOR 5G NETWORK RESOURCE MAPPING

5.1 Introduction

As already discussed in previous sections, in cellular networks, three mechanisms affect the appropriate placement of end devices to RATs/layers, namely: cell (re) selection, call admission control and handover. Cell (re)selection is a device control operation while call admission control and handover (in the case of horizontal handovers) are network controlled operations assisted by end devices. Due to their importance, these mechanisms have drawn significantly the attention of the research community.

During the past years, standardization bodies like 3GPP have specified well established procedures that typically employ simple algorithms (e.g., an end device or an eNB evaluates the received signal strength) in order to reach a decision. Researchers worldwide have built on these solutions and produced a rich set of algorithms both in scientific publications and patents. In order to provide more sophisticated mechanisms, the proposed approaches take into consideration additional parameters like the location and speed of a terminal, the experienced interference, the executed service, the required Quality of Service (QoS), the available bandwidth, the energy consumption, the user profile etc. All this information is referred as “contextual information” and can be used to improve network performance and eventually the QoE for the users. The main goal of all these mechanisms is to employ appropriate tools (e.g., utility functions, fuzzy logic, etc.) so as to evaluate the context information and reach a decision which optimizes the placement of the users in RATs and layers in terms of throughput/latency/delay or other KPIs.

An alternative approach is to monitor the behavior of a user (e.g., location, mobility pattern, use of specific services), analyze it and try to create a classification of end devices based on the respective user behavior. This classification can then be exploited by the mechanisms that affect the placement of end devices to RATs/layers. For example, there are users that use their smartphones only for placing phone calls, while others are “demanding” data users (e.g., web surfing, emails, games). For the first case, the network could place devices on legacy systems like for example GSM while for the latter they could be placed on LTE or Wi-Fi access networks. The innovation of this idea is to build a user profile on an automated way, by analyzing in offline mode a number of user-related parameters and combine them with available contextual information.

This novel proposal’s directions is also followed by 3GPP’s latest standardization moves, proposing the Network Data Analytics module (NWDA) ([154]). NWDA Function (NWDAF) provides slice-specific network data analytics to the PCF (Policy Control Function). NWDAF provides network data analytics to PCF on a network slice level and the NWDAF is not required to be aware of the current subscribers using the slice. The data may be aggregated from various network elements and functions, such as PCF (Policy Control Function), ANDSF, OFCS, NMS, etc.

The aforementioned concept is implemented by CEPE, a Context Extraction and Profiling Engine ([195]) and is analysed thoroughly in the following sections.

5.2 Knowledge Discovery (KDD) Tools

Methodologies and tools for KDD are divided into three categories: unsupervised, supervised and semi-supervised. *Unsupervised* KDD methods assume the existence of pattern(s) in data, which they try to unveil (e.g. identify *clusters* of similar observations). *Supervised* KDD methods on the other hand focus on learning existing pattern(s) from available data (i.e. training set) and then use them in order to classify previously unknown

observations (e.g. assign a new object to a set of predefined *classes*). Finally, *Semi-Supervised* KDD bridges the two aforementioned genres by attempting to identify pattern(s) in datasets (like Unsupervised KDD) using information provided by a limited training set. Before delving into the details of our framework we present a number of KDD algorithms which we employ in our work.

k-Means clustering

k-Means [196] is a partitioning clustering algorithm used for many unsupervised learning tasks. k-means tries to separate samples in k groups of equal variance *by minimizing the sum of intra-cluster distances* (see objective function below).

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - m_i||^2$$

where x is an object from the dataset and m_i the centroid of cluster C_i. k-Means is often referred to as Lloyd's algorithm ([196]).

Spectral clustering

Spectral clustering operates on the first k eigenvectors derived from the decomposition of the data graph Laplacian and applies k-Means on the projected dataset in order to derive the clusters. The algorithm operates on a large graph defined by the data similarities which may be a full graph, a k-nn graph (only the k-nearest neighbors of each object are retained) or an e-neighbourhood graph (only the points within range e are retained).

Naïve Bayes

Bayesian methods are a set of supervised learning algorithms based on Bayes' theorem ([196]). Bayesian methods are called "Naive" due to the fact that they operate under the salient assumption that class membership depends on only one variable rather than their combination. From a practical perspective, calculations are fast and simple; given a training set with C classes, we compute $P(C_j)$, $j=1...C$ for all classes of the set. Then, by considering the "naïve" approach we quantify $P(x_i|C_j)$ for all instances of class C_j and all values of variable x_j from the set of observations.

ID3 Decision Tree

Decision Trees are powerful learning mechanisms used for classification and prediction. The ID3 algorithm ([196]) is the most known and widely used due to its simplicity and effectiveness. ID3 iterates over the dataset and divides it along variables taking into account their entropy. Specifically, on each iteration, it goes through remaining variables, calculates their entropy and uses the variable with the smallest entropy in order to split the dataset. The algorithm continues on the remaining attributes until either all variables are used or there are no more observations to divide (e.g. empty dataset or all remaining data belong to the same class).

5.3 Overview of the proposed solution

The comprehensive overview of the various solutions proposed by the researcher community, leads us to the conclusion that most approaches are either too simple to implement but achieve sub-optimal solutions, or provide significant improvements but their complexity or the burden placed on the network components renders them unattractive for a real deployment by the operators. It is therefore evident that existing solutions need to cover a larger gap in order for RRM mechanisms to be able to efficiently and realistically support the real needs and requirements of 5G networks, with one of the primary challenges being to deal with the constantly increasing number of mobile users and bandwidth-intensive services via effective and efficient network planning.

A critical issue is related to the additional information that needs to be exploited by the network. A novel, overarching framework, on top of all three control schemes (i.e., cell (re)selection, call admission control and handover) should be able to take advantage of multiple sources of information (UE- or network- oriented) and extract from it additional knowledge. The new scheme that we envisage collects information about users, services and applications, terminals and network conditions and –based on offline processing and knowledge extraction– categorizes the UEs according to their behavior. From an architectural perspective, the derived models are provided to the Home Subscriber Server (HSS) and to the serving MME of a terminal so as to be exploited during the cell selection, CAC and handover processes.

Figure 56 describes the methodology for the knowledge extraction and the subsequent enhancement of the afore-described mobility control schemes. The first step is to identify the data that should be monitored, collected and processed as well as the Key Performance Indicators (KPIs) which will be used in the end in order to assess the effectiveness and efficiency of the derived model. **At least four data types and associated KPIs should be selected and used, namely: *network operation data, user behavior information, terminal capabilities and consumed service/application data*.** Afterwards, we proceed with the evaluation of the contextual information, which entails the extraction of data (e.g. selection of measurement according to location, time, etc.) and the derivation a model, which will be used in order to create enhanced cell (re)selection, CAC and handover mechanisms. The model is finally evaluated against the initially defined set of KPIs.

It should be noted that **CEPE is not an algorithmic solution but a KDD framework** focusing on the exploitation of available contextual information in order to dynamically identify profiles and associate them with sets of rules, which upon application can ameliorate the overall network operation (i.e., provide a more efficient RAT/layer mapping of UEs).

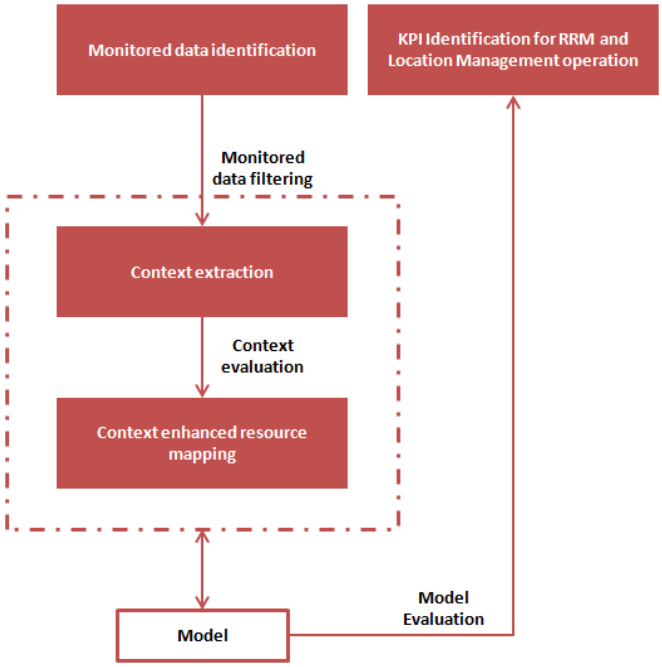


Figure 56: CEPE high-level methodology

At this point, we introduce the data-modeling step. Data modeling is a fundamental task of KDD since it formalizes the basic entities and parameters of the system under consideration. Therefore, it is an attempt to formally capture dependencies and

interaction between involved entities. Towards this end, we drafted a first data model, which set the basis for our work. We used in order to depict basic entities, their parameters and the interactions among them. In order to facilitate the KDD process and highlight the added value of our approach we included a limited set of parameters, essentially the minimum set of parameters that will help us draw conclusions from our data. Of course the model itself can be extended in order to accommodate current and future industrial data. The model is presented in Figure 57.

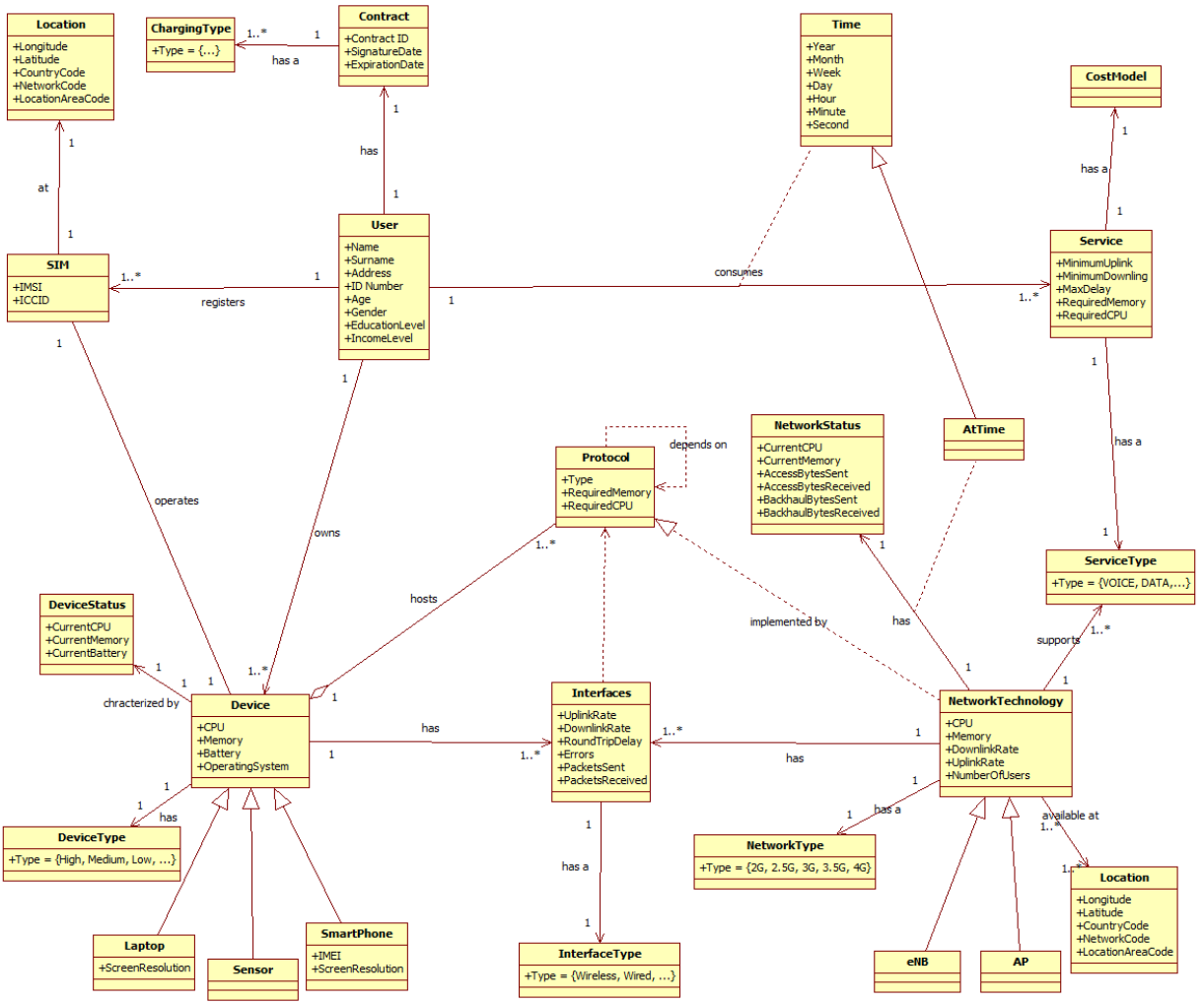


Figure 57: CEPE Data Model

The basic entities of the diagram are the *User*, the *Service*, the *Device* and the *Network Technology*. The user is characterized by a unique ID Number and a number of static attributes (e.g. Name, Surname and Address). The *User* may own one or more SIM cards and one or more mobile Devices. The SIM card is described by a number of attributes; IMSI and ICCID uniquely identify the card worldwide. In parallel, the SIM card holds the geo-location of a user described by entity Location. A *User* has signed one or more Contracts with one or more mobile operators. This Contract is described by a number of static details (date of signature, duration, annulment cases, etc.) and contains the charging details. The location is stored in two different formats, either using coordinates or by exploiting the Location Area Identifier (LAI). The *Device* is described a core set of static characteristics like Type, CPU, Memory, Batter, Operating System and has a number of subclasses that inherit and extend it. Depending on the inheritance level, a Device instance can be Smart Phone class (having IMEI and Screen Resolution as additional features), Laptop or a Sensor. A *Device* can hold one SIM card and has a dynamic state at specific time intervals described by CPU, Memory and Battery.

A *Device* may host one or more Protocols. A Protocol may be composed of other Protocols, thus realizing the concept of the protocol stack. The stack is a fundamental concept that enables the realization of an Interface. A device can have one or more physical Interfaces which are enabled due to the existence of proper Protocols. Each Interface has a Type and a number of parameters. A *Network Technology* is implemented through one or more interfaces and implemented by a set of Protocols. Network Technology is an abstract term encapsulating devices like APs, Routers, eNBs etc. Thus, it can a similar set of characteristics as the Device and is deployed in a certain Location.

A *Network technology* offers one or more *Service Types*, which essentially characterize Services. The latter are described by a Cost Model and a set of minimum requirements which must be met in order for the service to be offered through a *Network Technology* to the *Device* of a *User*. Of course, Service consumption takes place at a specific Time, thus every relation of our model is directly or indirectly characterized by the Time concept (all variables are populated with values at specific sampling intervals).

For simplicity reasons, we assume that each observation derived from our ecosystem resembles a row in a log file consisting of all the monitored parameters. For example, assume that a *User* with his *Device* starts at *Time* t_0 consuming a *Service* through a *Network* technology. *User* refers to the unique identifier of the specific UE, *Device* to the specific type of equipment that is used along with the device capabilities (CPU, monitor, etc.) and the *Service* relates to the type of the session that is active when the particular information is logged (i.e., VoIP call, browsing, Video streaming, etc.). Finally, the *Network* refers to the type of the RAT, as well as the cell layer via which, the specific service is being consumed (e.g., Wi-Fi, LTE femto-cell, macro cell, etc.). Consider now that we take a "snapshot" of the system every *t* seconds. Thus, our log input will be made up of numerous rows that look like:

$$\boldsymbol{User} \bowtie \boldsymbol{Device} \bowtie \boldsymbol{Service} \bowtie \boldsymbol{Network\ @Time}$$

Obviously, using the available information, we can derive additional parameters and augment our model. For example:

- *Uplink Peak User Throughput:* find the max value within a specific time frame
- *Downlink Peak User Throughput:* find the max value within a specific time frame
- *Uplink Delay:* calculate uplink delay within a specific time frame
- *Downlink Delay:* calculate downlink delay within a specific time frame

Similarly, we can quantify any required indicator and augment our input data:

$$\{User \bowtie Device \bowtie Service \bowtie Network\ @Time,\ KPI_1,\ \ldots,\ KPI_n\}$$

Note here that the level of granularity can change by selecting different time periods or a different entity. For example, we may ignore the *User* and *Device* axes of our data model while aggregate records per Week and evaluate the effect of *Service* consumption on our *Network*:

$$\{Service \bowtie Network\ @\ Time_{Week},\ KPI_1,\ KPI_2,\ \ldots.,\ KPI_n\}$$

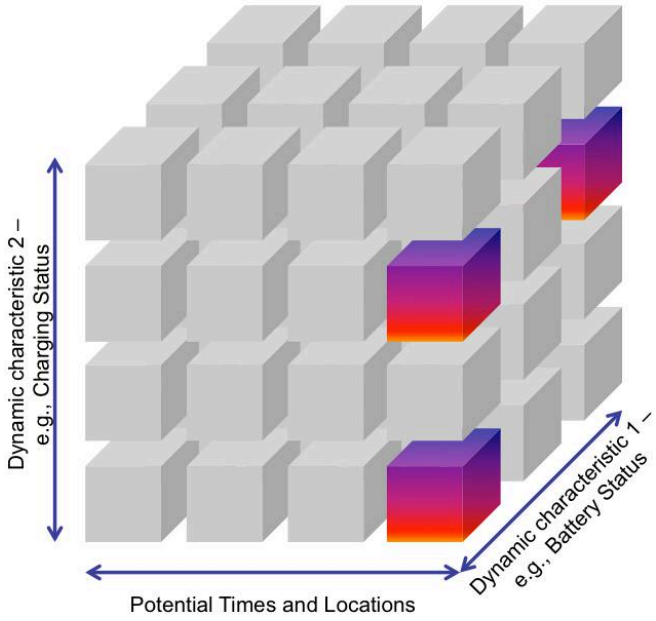


Figure 58: Graphical representation of the multi-dimensional profile

During the design of CEPE we took into consideration the huge amount of available information, as well as the need to formulate and derive various groups and associations, either apparent or latent, offline or at near real time. **CEPE therefore operates on two levels** (i) ***Macroscopic***/ Horizontal where each observation is treated as a whole and (ii) ***Microscopic***/ Vertical where all operations are applied on the elements defining each observation[6]. This two-level approach facilitates the application of CEPE on large data collections while in parallel –since a significant portion of the information remains static (e.g. device type) – speeds up classification. From a high level, methodological perspective, the proposed framework builds a is implemented in three steps:

- Data is aggregated from the identified data sources and formulates the dataset to be processed (*Level 0*)
- The dataset is broken down into subsets in order to derive **sets of similar observations** (*Level 1*, e.g. similar time period, similar geographic area, etc.).
- Each subset is broken down into entity specific chunks following a simple disaggregation approach; **each observation is broken down into its constituting entities** (*Level 2*, e.g. *User*, *Device*, *Network* and *Service* sets).

Similar observations per entity are grouped together (e.g. $\boldsymbol{User} = \{User_1, User_2, \ldots, User_N\}$). We assume that each group defines a node in a graph which is connected with another entity node via an edge with weight $w_{i,j}$ where *i* and *j* denote groups of different entities (e.g. $User_i$ and $Service_j$ are connected with an edge of weigh $w_{i,j}$).

The next step is to answer questions; for example, how to optimally assign a *User* to a RAT (e.g. cell ID and location) taking into account contextual information. A naïve usage of the model for this purpose would be consider the *User* entity, identify the *User*'s group, traverse the graph and find the most proper *User-Service-Device-Network* path (e.g. a path that maximizes the sum of weights). As we will see in the following paragraphs, we exploit this graph-traversal approach but apply a more elaborate scheme.

[6] Recall that an observation comprises the concatenation of instantiations of different participating entities (e.g. User x Device x Service Status @ Time)

Building and updating this model takes place offline, since it is time consuming. The updates are based on a sliding window approach where a set of observations is replaced by a new one. This action is executed periodically, in the data-warehouse of the network operator, since it implies rebuilding the models.

Searching on the other hand can take place in real-time. *Network* and *Device* nodes are few (devices can be roughly categorized into maximum 10 groups; types of networks even less) while user nodes are also limited by the input *User*. *The added value of CEPE is that it is generic in the sense that it can be applied to the whole dataset, a subset of the dataset or a time-projection of it. This means that you can obtain different models for different geographic areas, time slots etc.*

CEPE can be applied in either a supervised or an unsupervised context, thus leading to different results. In the unsupervised CEPE, we assume that there are some groups in our data, which we attempt to identify. On the contrary, in the supervised CEPE, we know in advance the groups (i.e., classes) and attempt to define a model that best describes them so as to be able in the future to categorize previously unobserved instances.

Evidently, everything comes at a cost; ***in the unsupervised case we gain flexibility and adaptability but the approach is prone to ill-defined and noisy data. In the supervised case we gain robustness but lose adaptability.*** For example, consider the case where a network operator identifies a new user group based on accumulated charging data records (CDRs) and customer information (CRM). Using this information, he sets up a campaign (e.g. offers free MBs to low spenders), which in turn results in the definition of a new group (e.g. low spenders who exhibit high-spender characteristics for a certain amount of time). Supervised CEPE will successfully categorize users in the initial group set but will fail to identify the new one. The unsupervised CEPE will exhibit worse performance in the first case –identification of profiles– but will find out that a new group has emerged when sufficient data is accumulated.

All the details related to the Supervised and Unsupervised versions' modelling of CEPE, along with the Rules Extraction KDD methodology can be found in the Annex. A final remark that needs to be made is that, besides the pre-processing stages, which take place in the context of the CEPE KDD methodology –and were previously discussed-, a parallel pre-processing of context information takes place: In the following section, CIP is presented, a context information pre-processing module, which acts in a distributed manner from the side of the different context-providing network entities, and minimizes the aggregated context information to be provided to CEPE, before even transmission. The direction of such a scheme is to minimize the network signaling overhead posed by such context aggregation mechanisms; on the contrary, the CEPE KDD part pre-processing module focuses primarily on the optimisation of the already acquired data towards the acceleration of the overall process.

5.4 CEPE deployment in the EPC network architecture

In this sub-section we present a graphical representation of how this scheme is deployed in the actual EPC network architecture (Figure 59).

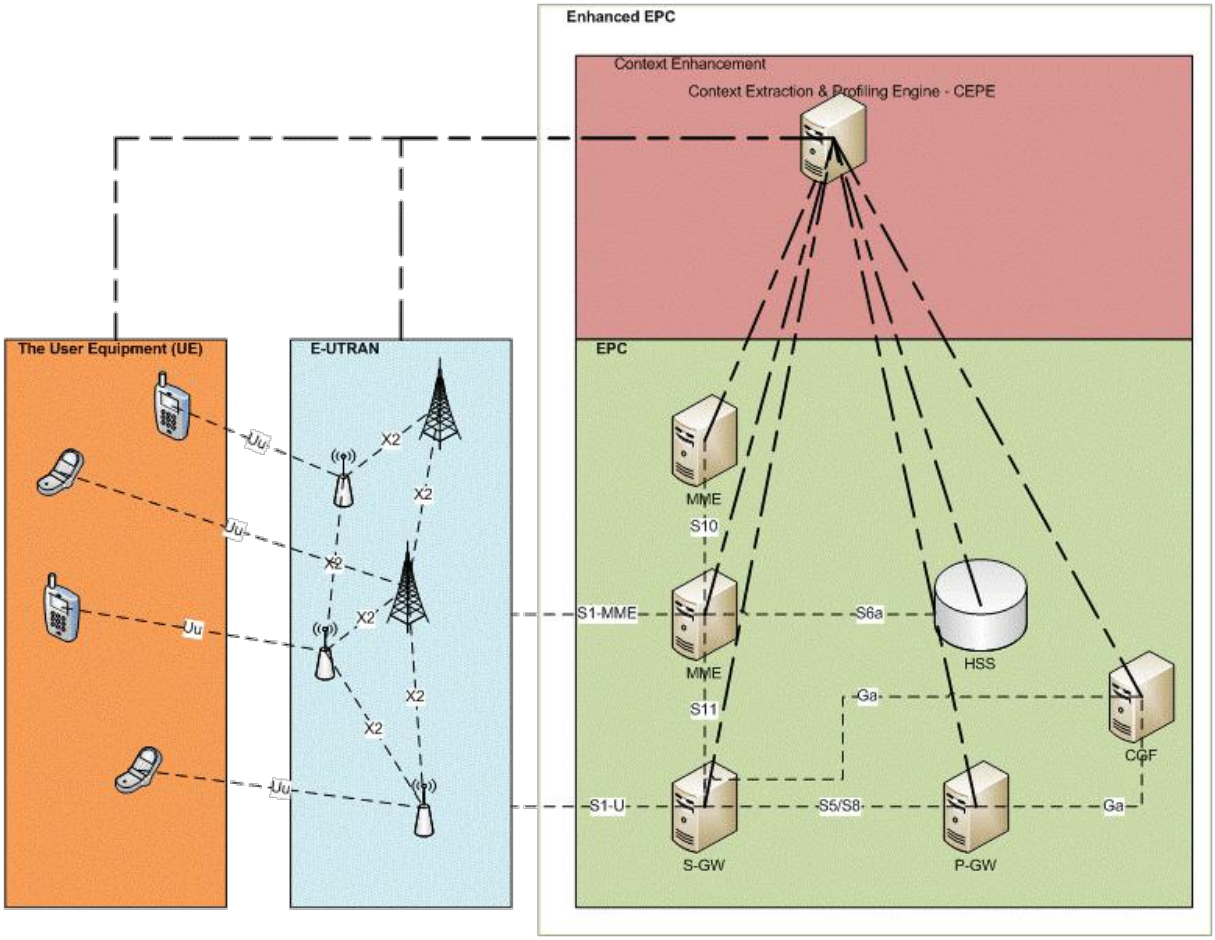


Figure 59: CEPE deployment in EPC network

As already discussed, CEPE collects information about users, services, end terminals and network conditions and executes the offline processing in order to categorize a user according his behaviour. The resulted information is then passed to the HSS and to the serving MME of a terminal so as to use this information eventually by the network components during the cell (re)selection, the CAC and the handover processes. The deployment of CEPE requires obviously new interfaces in the network architecture.

The proposed solution encompasses all the key functionalities, namely, Context Acquisition, Context Modelling, Context Exchange, and Business Logic. More specifically, the proposed CEPE scheme incorporates, information gathering from several network points (e.g., UEs, (H)eNBs, HSS, PGW, SGW, MME, etc.), which is the Context Acquisition and Context Exchange functionalities; the correlations among the inputs are being identified afterwards (Context Modelling) and then are being linked to specific business goals automatically (Business Logic).

Figure 60 depicts the CEPE operation, highlighting the introduction of CEPE entity, the interfaces which are being considering for extensions as well as exemplary information exchange. In this, CEPE gathers user information from HSS, service information from AS (Application Server), UE capabilities and status from the UEs, and network resources information from the network entities involved (e.g. eNB). After the knowledge extraction process has been completed, CEPE communicates the device classification outcome to the network entities and the UE.

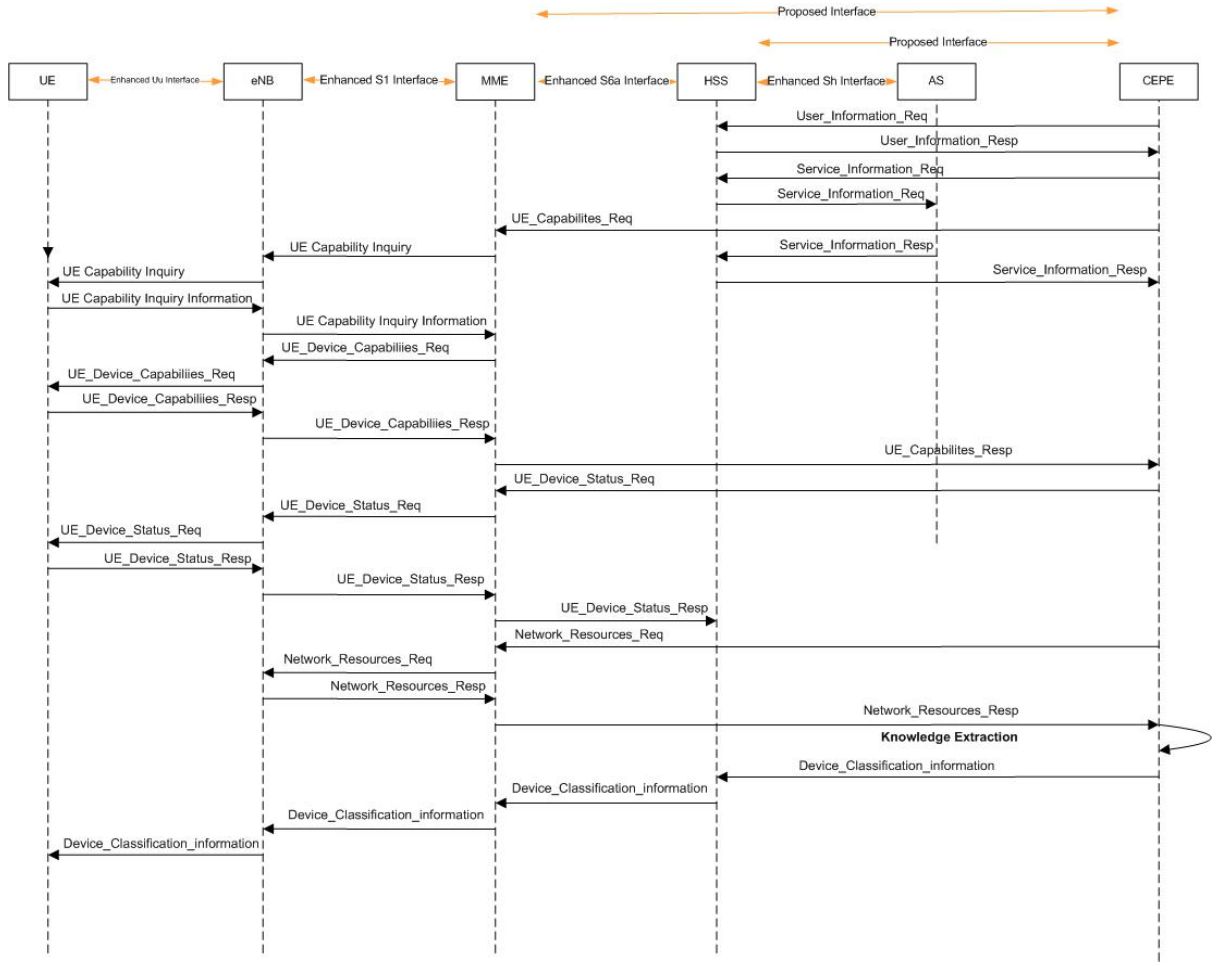


Figure 60: CEPE operation example

5.5 CEPE Experimental Evaluation

In this section, in order to assess the validity and viability of the proposed framework, CEPE is evaluated via the realization of 2 different experiments. Through the experiments we attempted to replicate –to the best possible extent– a real life situation. Towards this end we conducted an extensive literature review that covered a large number of aspects like mobility speed, energy consumption patterns, service usage patterns, etc. We present our findings and configuration in the following.

5.5.1 CEPE 1st Experiment

Experimentation Scenario and Setup

This experiment considers a Shopping Mall environment, as this was described in the evaluation of COmpAsS mechanism (see Experiment 2: Realistic Shopping Mall use case), i.e., an established 5G scenario, part of the UDN use case.

A typical setting for a future extended rich communication environment, involves “traditional” radio networks and wireless sensor networks, where customers access mobile broadband communication services while they are directly addressed by personalized location-based services of the shopping environment. Overall, the network deployment allows seamless handling of services across different domains, e.g. mobile/fixed network operators, real estate/shop owners, application providers. Based on this description, we use the NS3 and model a single floor, 200x100m building, containing 10 rooms, with an LTE Femto cell placed in each of them. Outside, two LTE eNBs are

placed, 150m north of the mall with Inter-Site Distance (ISD) equal to 200m, and a GSM cell between them (at equal distance from the eNBs).

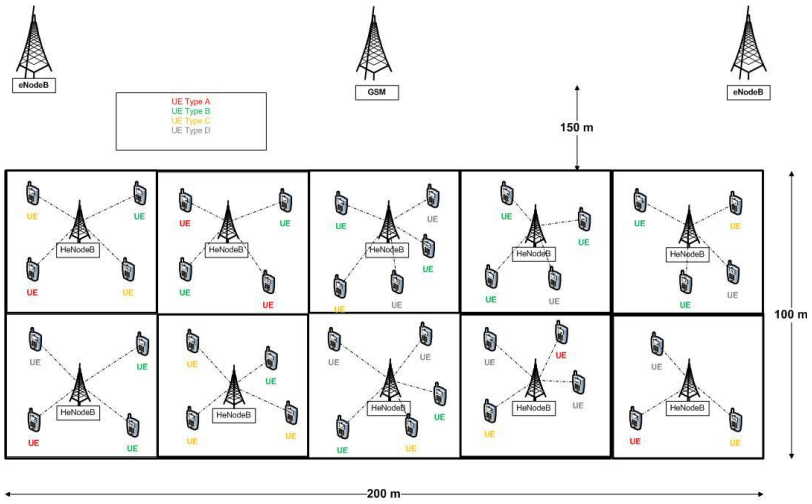


Figure 61: CEPE 1st experiment simulation topology

Our simulation scenario is based on 3GPP Specifications [185] and [186]. In details, the transmission mode is SISO (Single Input Single Output); the handover algorithm is the A2A4 RSRQ-based7 and the scheduler is the NS-3 implementation of the Proportional Fair MAC scheduler [186]. We use the Hybrid Buildings Propagation Loss Model for path loss implemented in NS3 with Internal Wall Loss at 10.0 dB Shadow, Sigma Indoor at 10.0 dB [185]. The network node configuration appears in Table 15. Services are implemented using NS3's UDP client-server application model and the desired data rates are achieved through configuration of the packet size and the inter-packet interval parameters. The service schedule for every user is pseudo-randomly generated at the beginning; as the simulation progresses they affect and are affected by the battery state and the charging status. Service parameters appear in Table 16: Each time one of the services is triggered according to the service schedule mentioned above, a constant bit rate traffic model is generated with the respective duration; the traffic is between the clients (UEs) and a remote host, while our measurements concern only the part of the access network.

Every user follows a mobility model comprising a) the velocity and b) the path pattern (linear, random, etc.). The mobility model may change during simulation. Every 4 minutes, each user randomly selects a model; additionally, when a UE has moved 30 meters towards any direction, it randomly selects another direction to move next. The considered mobility models are: Stationary Mobility (0 m/s – 0.8 m/s); where customers move very slow or remain at their position, Low Mobility (0.8 m/s – 1.4 m/s); where customers move with a slow or average pace inside the mall and Medium Mobility (2 m/s +/- 0.6); where customers walk fast inside the mall. Moreover, every user has a charging level denoted as Bronze, Silver or Gold (randomly assigned based on a uniform distribution), emulating the data capacity of his subscription. Bronze users have a maximum of 500 MB to spend on data services (the initial value is randomly generated between 40-500 MB), Silver users a maximum of 2GB (the initial value is randomly generated between 100-2000 MB), and Gold users have no threshold.

[7] https://www.nsnam.org/docs/models/html/lte-design.html#fig-lte-legacy-handover-algorithm

Table 15: NS-3 configuration

NS3 Network Node	Tx Power (dBm) [185]	Downlink (DL) Earfcn (MHz) [185]	Bandwidth (RBs) [185] [186]	Antenna Type [185]
Macro cell	35	2120	50 (10 MHz)	Parabolic, 15 dBi
Femto cell	20	2120	15 (3 MHz)	Isotropic
GSM	35	2120	15 (3 MHz)	Parabolic, 15 dBi
UE	20	-	-	Isotropic
Macro cell	35	2120	50 (10 MHz)	Parabolic, 15 dBi

When data services are used, the available data that a user has according to his subscription are reduced. Recall that in real life, users tend to reduce their activities (i.e. data usage and session duration) when their data availability becomes low. In order to replicate this behavior, we assume that when bronze users consume 80% of available data they reduce their activities to 10% of their normal habits and corresponding session duration to 70%. Similarly, silver users reduce their data intensive activities to 50% when they consume 80% of their available data and cut their duration by 50%.

Table 16: Service parameters used in simulation

Service		***Value***	***Comments***
Type	***Characteristic***		
Short Duration Voice	Duration	100 sec (+/- 10)	Average call duration is 1.8'[187] Average rate is 12.65 kbps[188][189]
	Rate	13 kbps UL & DL	
Long Duration Voice	Duration	240 sec (+/- 20)	
	Rate	13 kbps UL & DL	
Web Data	Duration	4 sec (+/- 2)	Average web page access session duration is 4.2 seconds [200] Average web page size is 1.6MBs [201][201]
	Rate	1.6 MB DL (+/- 500KB)	
FTP Data	Duration	9 sec (+/- 2)	Average session for file download is 9.8 seconds for 3MBs file [200]
	Rate	3 MB DL (+/- 60KB)	
Video Stream Data	Duration	240 sec (+/- 30)	Average YouTube video duration is 4.12' [192] Average DL speed 443 kbps [191] Average size for 480p video is 250MB per hour in YouTube[193]
	Rate	13.5 MB DL (+/- 1.5)	
VoIP Data	Duration	900 sec (+/- 300)	

	Rate	45 kbps UL & DL	Skype audio only UL 42-47 kbps, DL 42-47 kbps [191] Average Skype call duration is ~20 minutes [202]

We consider 3 different device classes, namely high, medium and low capabilities terminals, which affect the total battery capacity and battery reduction of a UE. The battery consumption formula is based on the battery's maximum capacity and the battery's consumption of each service on every device. In the case of smaller (Femto) cells, we consider that the consumption is proportional to the Macro cell's consumption, due to the fact that less transmission power is needed on the UE side. Throughout the simulation we consider that users do not change devices.

Battery status is initialized similarly to charging status (i.e. uniform distribution, but taking into account maximum capacity due to different devices). Furthermore, we consider 3 distinct battery levels, namely *High*, *Medium* and *Low*, each having an impact on the user behavior. Initial battery state is randomly distributed between 20% and 80% of maximum capacity. A device has High battery status when more than 35% of the total battery capacity is available. A high battery status has no effect in the rate or duration of the services used throughout the scenario. A Medium battery status is assigned to devices operating between 10% and 35% of total battery capacity. When a device's battery status drops to Medium, the user consumes 50% of the calls/services he would normally consume (i.e. in High battery status) and their duration is cut by 50%. Finally, a low battery status is assigned to devices operating below 10% of total battery capacity. When a device's battery status drops to Low, the user only uses 10% of calls/services he would normally consume (i.e. in High battery status) and their duration is reduced by 70% (duration reduction does not apply for web pages). Table 17 below describes in detail the characteristics of the three device classes and the battery models.

Table 17: Types of devices and associated battery consumption

Device Type	***Screen Type***	***Battery Capacity (mAh)***	***Macro Cell Battery Consumption***	***Femto Cell Battery Consumption***
High Capabilities Device (Based on Nexus 5)	LCD 1080*1920, 4.95inch	2300	145,27*t(talk) + 483,19*t(web) + 377,04*t(video) + 7,66*t(idle)	0.714 * Macro cell consumption
Medium Capabilities Device (Based on Samsung S3)	Super AMOLED 720*1980, 4.8inch	2100	215,38*t(talk) + 325,58*t(web) + 222,22*t(video) + 2,65*t(idle)	
Low Capabilities Device (Based on Nokia E66)	TFT 240*320, 2.4inch	1000	133,33*t(talk) + 230,77*t(web) + 312,5*t(video) + 2,97*t(idle)	

Experimentation Methodology

Using this scenario and assumptions we generate a number of datasets upon which CEPE was evaluated. During simulation time we monitor the status of the user, device, network and service and store it into a log file. Each record of the dataset contains the time of the observation, the UE's International Mobile Subscriber Identity (IMSI) and all of the information we can collect related to UE's, services' and network's state at the specific time of the observation. An excerpt of this information is depicted in the following figures (Figure 62Figure 63).

imsi	name	surname	age	gender	education	income	cpu	cores	os	screen_width	screen_height
1	Maria	Taylor	36	F	College	33220	1.3Ghz	dual	iOS	240	320
2	Nick	Lukas	19	M	not graduated Highschool	24712	1.3Ghz	dual	iOS	240	320
3	George	Brown	22	M	Higher than College	22394	1.7Ghz	dual	Android OS	720	1280
4	George	Smith	29	M	Higher than College	24916	1.6Ghz	quad	Android OS	1080	1920
5	Lusille	Brown	54	F	College	47098	1.6Ghz	quad	Android OS	1080	1920
6	George	Tremblay	58	M	Higher than College	56884	1.3Ghz	dual	iOS	240	320
7	Nick	Smith	78	M	College	16722	1.6Ghz	quad	Android OS	1080	1920
8	Constantine	Smith	50	M	not graduated Highschool	18461	Unknown	Unknown	Symbian	240	320
9	John	Smith	51	M	Highschool	23208	Unknown	Unknown	Symbian	240	320
10	Luisse	Skywalker	43	F	Higher than College	15122	1.3Ghz	dual	iOS	240	320
11	George	Skywalker	44	M	not graduated Highschool	23320	1.7Ghz	dual	Android OS	720	1280
12	Nick	Skywalker	28	M	Higher than College	181582	1.3Ghz	dual	iOS	240	320
13	George	Wong	30	M	College	77579	1.3Ghz	dual	iOS	240	320
14	Martha	Brown	21	F	Higher than College	5488	1.7Ghz	dual	Android OS	720	1280
15	Maria	Brown	50	F	College	30939	1.6Ghz	quad	Android OS	1080	1920

Figure 62: Sample of the user and device static characteristics

time	imsi	x	y	z	velocity (x:y:z)	txpower	ultx	sum_ultx	dlrx	sum_dlrx	ul_delay	dl_delay	lost_packets	associated_cellid	rsrp	rsrq
375	5	18.1683	25.2491	9.08511	00:00:00	20	0	4210	18	17364	0	0.111069	0	2	-79.0004	-6.0066
375	21	52.2987	68.3325	11.1514	-1.95099:-1.56354:0	20	25	11359	26	20696	0.0123813	0.00714496	0	1	-77.2772	-5.48997
375	22	69.5819	24.7065	4.79334	0.770973:1.97211:0	20	0	10944	0	25848	0	0	0	8	-78.8653	-8.30015
375	23	73.6172	76.5321	4.62715	-2.20642:1.29282:0	20	25	10720	25	24558	0.0122713	0.00334262	0	1	-77.2884	-6.45531
375	24	45.9436	7.94052	4.61572	1.0997:-2.01332:0	20	25	9964	25	23387	0.0126559	0.00372722	0	6	-66.9001	-4.2853
375	25	38.2861	90.6926	11.9442	-2.54065:0.860013:0	20	0	13824	19	25391	0	0.554631	0	1	-76.0733	-6.23794
375	26	30.7951	40.5527	12.9944	2.62772:-0.491851:0	20	0	12020	18	25496	0	0.552982	0	1	-78.2644	-5.32413
375	27	70.3948	28.1009	6.23434	3.48727:1.06985:0	20	25	11327	15	25341	0.0124184	0.109823	0	8	-76.8916	-8.68234
375	28	140.162	49.6648	11.8284	-0.034281:3.55078:0	20	25	11309	25	23303	0.012905	0.0244163	0	1	-76.3161	-6.0656
375	29	100.556	51.5388	6.39348	0.930197:-2.26329:0	20	0	10673	0	22124	0	0	0	11	-78.6731	-5.80526
375	30	171.34	26.2544	2.14341	3.83806:-0.799411:0	20	0	11034	0	23307	0	0	0	16	-75.897	-7.50523
375	31	170.5	87.0532	5.88935	-1.42696:-3.05206:0	20	0	6393	0	20687	0	0	0	1	-66.8345	-4.11972
375	32	14.1901	78.9689	7.97035	-0.515784:2.77083:0	20	25	9791	26	20536	0.0124892	0.00613741	0	1	-76.4907	-5.66534
375	33	186.025	39.6246	9.94747	-0.446031:-2.35723:0	20	0	8065	0	17411	0	0	0	1	-80.7119	-6.22565
375	34	194.926	67.5309	3.0497	0.546728:-2.56947:0	20	0	8337	11	21214	0	0.154213	0	21	-69.1081	-4.34456
375	35	83.7675	10.6315	7.6823	2.71138:0.0945869:0	20	0	9740	0	22202	0	0	0	8	-71.2933	-6.08535
375	36	144.869	19.5153	12.0044	2.99819:1.88492:0	20	0	11859	0	22371	0	0	0	16	-75.2858	-7.65617
375	37	91.7005	90.1601	11.3707	-3.11177:1.98614:0	20	0	10609	0	19951	0	0	0	1	-77.1156	-7.36175
375	38	33.7284	98.4918	0.251241	0.824956:-3.56771:0	20	0	7209	0	23529	0	0	0	5	-70.8778	-5.85312
375	39	50.1913	89.3476	14.7854	-1.99062:1.13374:0	20	0	9910	24	20846	0	0.488858	0	1	-76.6427	-6.00061
375	40	67.6137	3.07486	6.3801	-3.118:1.33938:0	20	0	7128	0	20673	0	0	0	1	-70.8111	-5.38928

Figure 63: Sample of user and device dynamic measurements.

Data are post-processed in order to add labeling information; every instance is attributed a label based on the scenario assumptions of the previous paragraph. We consider this labeling information as the ground truth, i.e. the correct labels that our model should identify. Moreover, we map continuous variables like income and age to nominal values and string data to integer values (e.g. an education level 'College' is mapped to 2). This way we manage to map all variable values to real numbers and thus map any instance as a high dimensional point residing in R^n, where n is the number of variables describing each observation.

The final dataset is provided to CEPE for training. As soon as the knowledge base is built we rerun the same experiment using the derived rule-set and the learnt model. The second time, NS-3 uses the CEPE model in order to:

i. classify a user according to his behavior and

ii. identify the best set of actions to apply given the results of i.

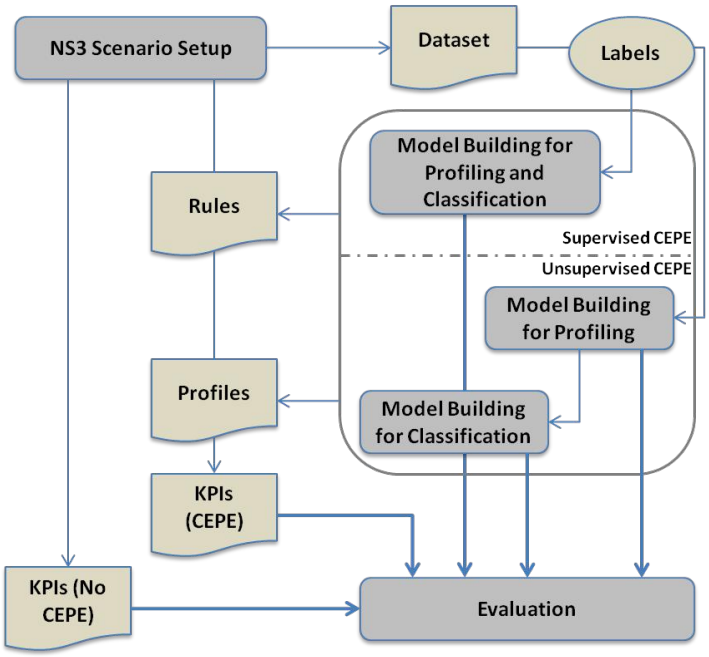


Figure 64: Experimentation methodology

Evaluation is performed along two axes; at first we assess the classification capability of CEPE itself and then we assess its effect on network conditions.

Both supervised and unsupervised CEPE versions are evaluated with respect to their ability to build a model that identifies new observations. In the unsupervised CEPE case we assess

i. the ability of Spectral Clustering to build a model that is close to the original labeling scheme

ii. and the ability of ID3 to properly model it and classify new instances with it.

The evaluation of the supervised case is simpler since we only assess the ability of CEPE to build a model which is close to real data.

In order to assess the quality of the clustering and classification algorithms we employ the F-measure:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

and

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

We retain the best performing model and proceed along the second evaluation axis that focuses on the effect of CEPE rules and profiles on network conditions.

We quantify mainly two KPIs, namely: *Uplink (UL)/Downlink (DL) Throughput* (i.e. the total throughput for the uplink and downlink respectively, for the entire simulation) and *Number of Handovers* (i.e. the total handovers performed during the entire simulation). The latter, i.e. *Number of Handovers*, is directly linked to QoS: the more handovers are realized, the higher the degradation of the respective on-going service performance, due to the handover signaling overhead increase, data lost during the handover duration, re-

transmission timeouts, etc. The overall experimentation methodology is depicted in Figure 64.

We employed two simulation models: *Low Background Traffic*, where during experiments current traffic was augmented by a +5% UL and DL and *Medium Background Traffic*, where current traffic was augmented by +20% UL and DL. The duration of each simulation equals 20 real-life minutes and we logged measurements every second. We considered 40 users, using the discussed service types (5 types), mobility patterns (3 patterns), device types (3 types) and charging categories (3 categories). Every scenario was executed 5 times thus all reported values in the rest of the section comprise the mean results of these executions.

Results

The application of the supervised CEPE resulted in a model, which attains an F-measure of 0.984 (e.g. the derived rules from ID3 can correctly classify 98.4% of observations). The background traffic did not affect the classification ability of the models. The unsupervised CEPE was more laborious since it entailed the initial application of Spectral Clustering. Specifically, we used a fully connected graph and the Gaussian kernel with σ=38. Spectral clustering managed to attain an F-measure of 0.90 (e.g. identified the correct number of clusters and correctly classified 90% of the instances). The application of the decision trees on the labeling scheme of the spectral clustering produced an F-measure of 0.95. Table 18 provides the aforementioned results both holistically but also per entity (e.g. device entities only). Obviously, it is easier to build a good model for Devices but it gets more difficult when the device classes are combined with the various mobility types and services. The latter was the reason that laid us to present the dynamic user information (i.e. mobility data) separately.

Table 18: Unsupervised and Supervised CEPE results -Knowledge Discovery Capability assessment w.r.t F-measure.

CEPE KDD Evaluation		*Unsupervised*		*Supervised*
		Spectral Clustering	*Decision Trees*	*Decision Trees*
Low Background Traffic	*Device*	0.97	1	1
	Mobility	0.89	0.951	0.981
	User	0.95	1	1
	Service	0.91	0.934	0.973
	All	**0.90**	**0.95**	**0.984**
Medium Background Traffic	*Device*	0.97	1	1
	Mobility	0.89	0.951	0.981
	User	0.95	1	1
	Service	0.91	0.934	0.973
	All	**0.90**	**0.95**	**0.984**

The experiments showcase a clear superiority of the supervised case, constantly exhibiting a better F-measure. In fact, when the unsupervised CEPE was evaluated

[8] The derived affinity matrix offered the best discriminative depiction of the underlying data clusters (i.e. clusters were better separated and clearer compared to other configurations).

against the ground truth labeling scheme, the F-measure was marginally equal to 0.90 giving the supervised case a clear precedence of almost 10%.

It is also worth noting that in the supervised case, classification results and associated rules are derived faster. The latter is due to the complexity of the employed approach; spectral clustering necessitates $O(|\mathbf{X}|^2)$ memory and $O(|\mathbf{X}|^3)$ processing time contrary to $O(|\mathbf{X}|)$ memory and $O(n|\mathbf{X}|\log|\mathbf{X}|)$ time required by decision trees -n is the number of features.

This evaluation also indirectly assesses the capability of CEPE to identify behavioral changes and adapt (i.e. identify that a subscriber suddenly changes mobility pattern); this is captured in the overall results. Finally, the lower quality was anticipated due to the unsupervised nature of the algorithm. *Thus, in the remainder of the section we focus on supervised CEPE.*

Using the methodology about Rules Extraction and Feedback Loop in Annex, and exploiting the derived classes-graph we extracted a number of rules. We used the semi supervised approach where we extracted the full rule-set from the graph, ranked it according to the sum of weights and selected the rules, which we deemed more suitable for ameliorating the KPIs. The selection we performed was based on the following assumptions / considerations:

- Gold users should have the highest possible quality of experience.
- High mobility users shall be served by macro-cells to minimize the number of handovers.
- Calls of higher duration by moving users shall be served by macro-cells since the probability of a handover during the lifetime of the call is higher than for short calls.
- Voice Services shall be served by a second/third generation technology so that resources in 4G systems are allocated to high data users.
- Low-end devices that cannot support advanced services should be served by a Second or Third Generation technology (e.g., GSM).
- Low Battery level devices shall use short-range wireless technology.

These rules are used in conjunction with the A2A4 handover algorithm. Each rule selects the most appropriate RAT (e.g., FemtoCell), and A2A4 undertakes the selection of the best FemtoCell where a UE should be handed over, based on RSRQ.

Performance evaluation per RAT

Our first evaluation depicts the results in relation to the RAT of the UE. Essentially, for every employed KPI we attempt to quantify it on a holistic level (e.g. how the packet loss was affected by CEPE) as well as on RAT level (how the packet loss in GSM was affected by CEPE). Towards this end we provide Figure 65-Figure 68.

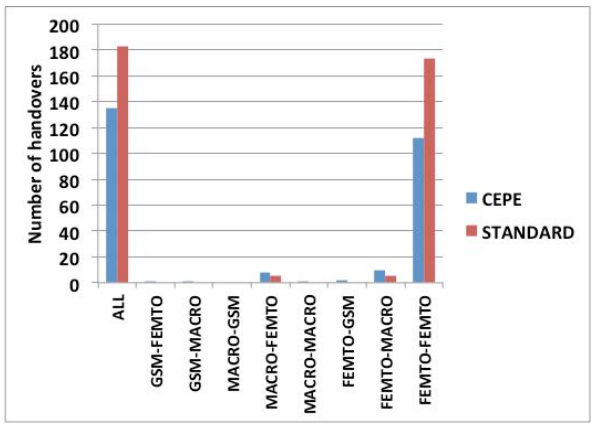


Figure 65: Number of handovers per RAT type - Low Traffic

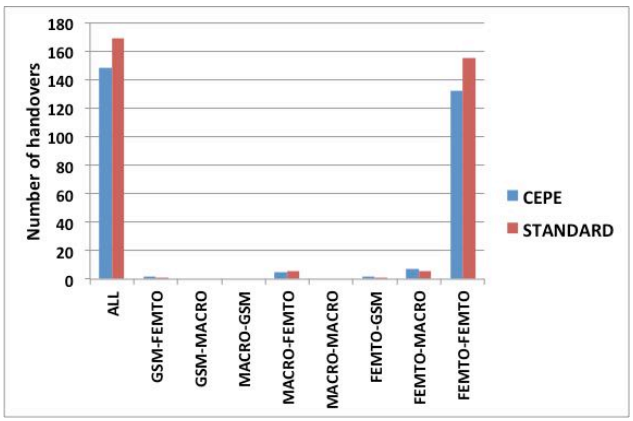


Figure 66: Number of handovers per RAT type - Medium Traffic

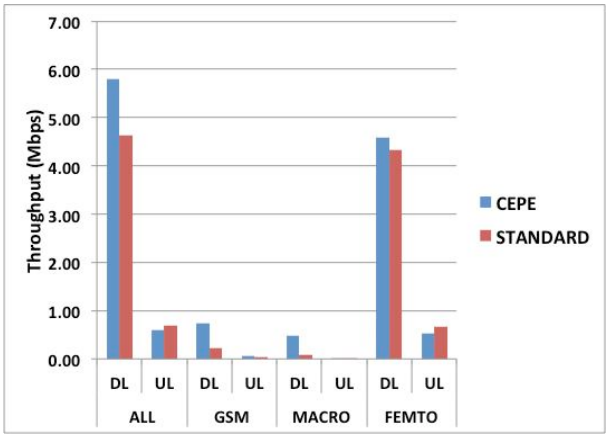


Figure 67: Experienced throughput per RAT - Low Traffic

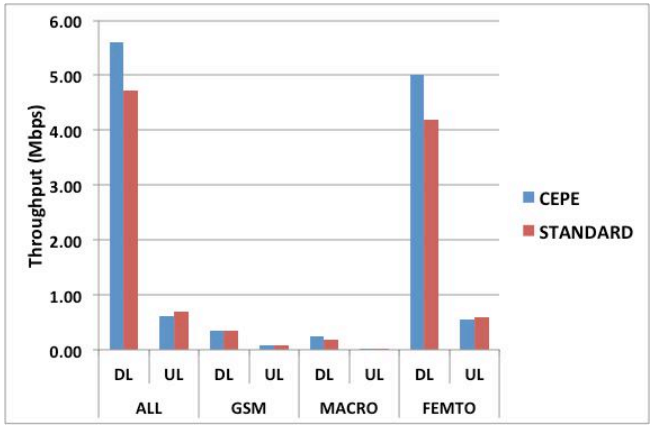


Figure 68: Experienced throughput per RAT - Medium Traffic

The four initial figures illustrate the overall number of handovers per RAT type, as well as the experienced throughput per RAT, both for medium and low background traffic. The graphs show that the application of CEPE reduces the overall number of handovers in both cases; in fact, the realized handovers are minimized by 15 – 20%. Looking closer, we observe that the femto-femto handover type dominates the overall picture. The reasoning for this fact is primarily related to the Ultra Dense Environment of the simulation topology that has been selected, and not the specific policy rules that were applied. There is a big number of femto cells inside the interior of the shopping mall, very close to which the UEs are moving. This results to high RSRP/RSRQ values, boosting as a result the femtos' selection.

Despite this fact, however, CEPE seems to considerably ameliorate the ping-pong effects and the often handovers – even in the very frequent event of femto to femto handover-; this is on the one hand due to the policy rule that high mobility users are never placed in femto cells (thus these users never participate in FEMTO-FEMTO handovers); on the other hand, the rule, which places the Voice Service UEs in the macro cells also decreases the possibility for UEs to need to realize a handover between adjacent FEMTO cells.

The throughput results are similar in both experimentation scenarios, implying that the performance of the CEPE-enabled handover scheme is not influenced by the traffic volume. From a high level perspective, the downlink throughput is improved by 500-1000 Kb/s. Throughput increases almost for all type of services, since CEPE takes into account the requirements for each service in order to allocate each user to the appropriate RAT. In the uplink case, the overall throughput is slightly reduced, mainly due to the reduction

of the femto cells' UL throughput. It must be noted, that the experienced throughput of the femto cell users is higher compared to other RATs, due to the fact that higher data rate services users (i.e., VoIP and Video users) are primarily handled by femto cells. In other words, there is a slight overload of the femto cells since all the data demanding services are handled by them while in parallel other type of RATs users experience lower throughput because they consume low data rate services.

It is worth mentioning the fact that we observed additional advantages in the ul/dl packet loss and delay (although we are not employing them as assessment KPIs). In the CEPE-enabled case the packet loss for the downlink, which accounts for ~90% of the traffic, was decreased by more than 4%. Moreover, a in the CEPE-enabled handover scheme the delay for the downlink was also decreased by almost 10%. Similarly to the throughput results' analysis, downlink delay was decreased for the LTE macro cells and femto cells since CEPE allocated each Service to the appropriate RAT considering QoS requirements and mobility type. The downlink delay was increased for the GSM BS (i.e. Voice calls) due to the increased distance and the different configuration of GSM comparing to other LTE technologies; still however, it is in the acceptable limits (i.e., below 200ms).

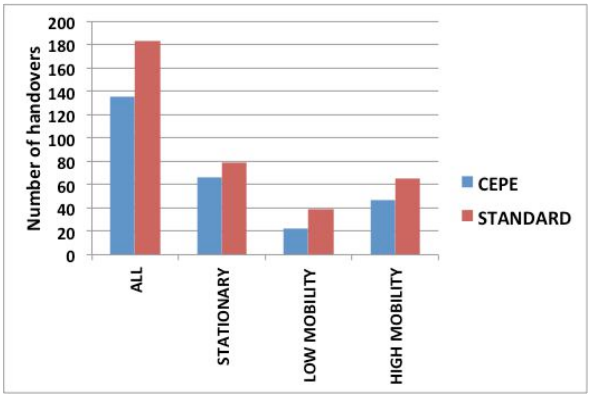


Figure 69: Number of handovers per mobility type- Low Traffic

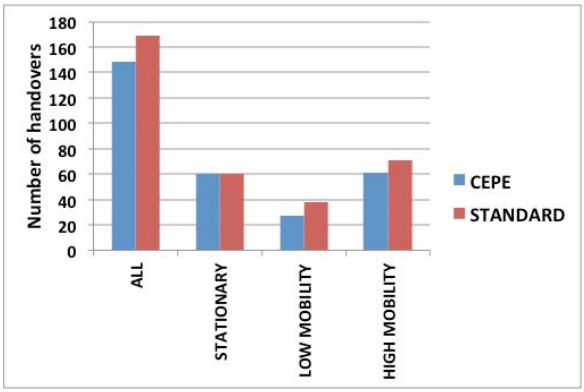


Figure 70: Number of handovers per mobility type - Medium Traffic

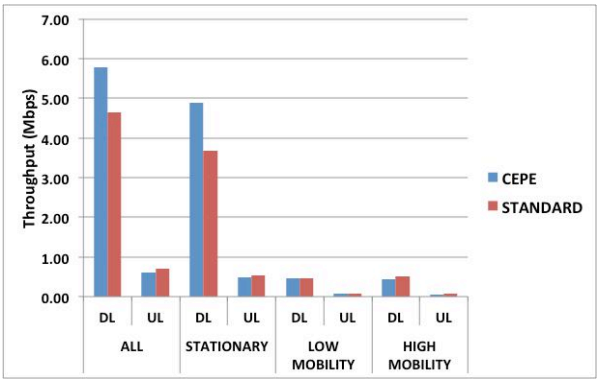


Figure 71: Experienced throughput per mobility type - Low Traffic

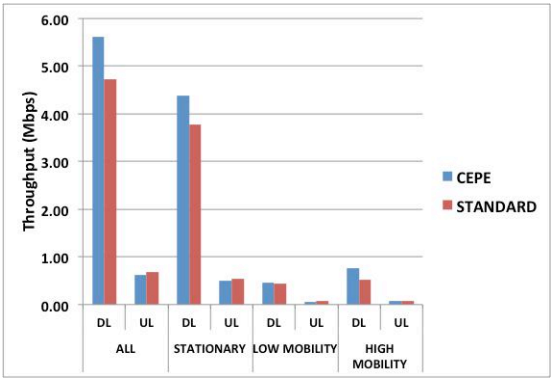


Figure 72: Fig 8: Experienced throughput per mobility type - Medium Traffic

Performance evaluation per Mobility Type

The second evaluation perspective illustrates the simulation outcomes in relation to the type of UEs mobility. Similarly to the previous paragraph, we provide all the results in Figure 69Figure 72, presenting the overall number of realized handovers and experienced throughput, both for low, as well as for medium traffic experiments.

The CEPE-enabled network manages to decrease the overall number of handovers in both scenarios and for all mobility type cases. Only the stationary users in the medium traffic scenario experience the same number of handovers. The minimization of the

overall handovers is achieved by the rules tendency to place each user to the optimal RAT according to the profile, which has been extracted and avoid consecutive handovers, based only on the RSRQ metric.

The experienced throughput in the CEPE-enabled handover scheme case is also improved in both traffic scenarios. The most significant gain is observed in the case of stationary users; in the low background traffic scenario the average gain for CEPE is equal to 33%, while in the medium background traffic it is slightly over 19%. Note that the CEPE-enabled scheme shows similar performance to the standard scheme in almost all other cases. In addition, we observe that the mobility is inversely correlated with the throughput gain; the lower the mobility the higher the throughput gains. This is mainly associated with the rule that switches all stationary users with high throughput Services (e.g., Web or FTP) to femto cells. In the case of uplink, a slight decrease (i.e., 2-6%, depending on the case) is observed in all mobility types, mainly due to the interference. The latter is the result of the fact that the number of users that are associated to a macro cell in the CEPE experiment is four times higher when comparing to the standard.

Performance evaluation per Service Type

During the simulations, diverse types of services were deployed in the UEs. The possible service types were discussed earlier. Similarly to the previous paragraphs, we provide all the results in Figure 73Figure 76.

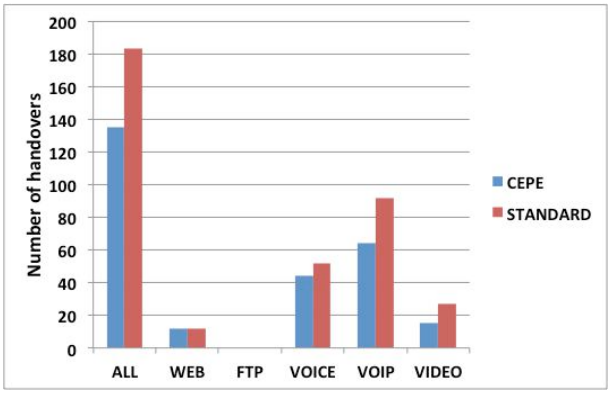


Figure 73: Number of handovers per service type- Low Traffic

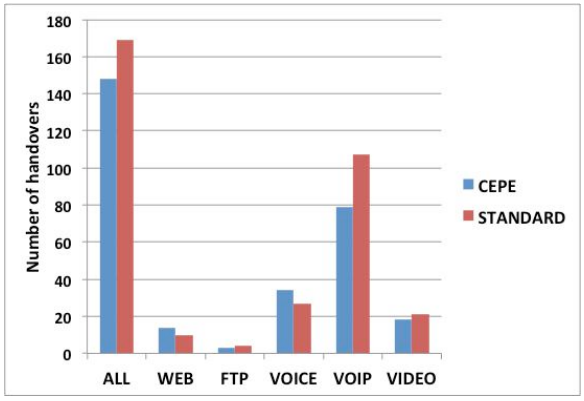


Figure 74: Number of handovers per service type - Medium Traffic

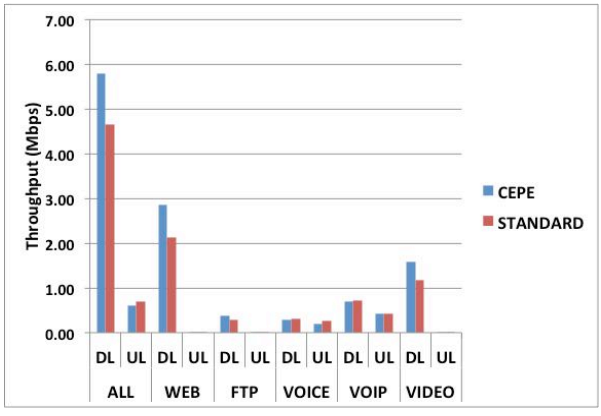


Figure 75: Experienced throughput per service type- Low Traffic

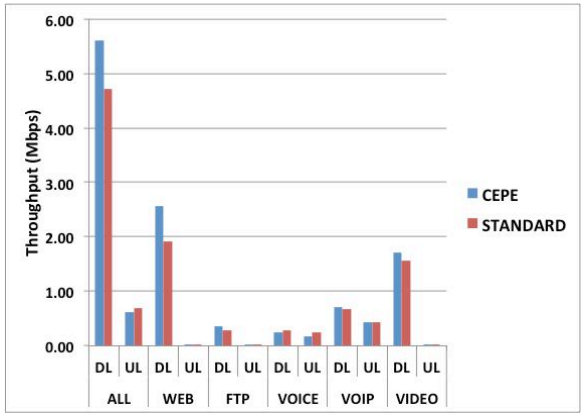


Figure 76: Experienced throughput per service type - Medium Traffic

The first KPI, which is illustrated in the figures, is the overall number of handovers that took place per Service type. The latter is decreased for all service types when deploying the CEPE-enabled handover scheme, apart from the Web type of service case, in which the CEPE scheme realizes equal number of handovers with the standard scheme. The CEPE-enabled handover mechanism shows an overall enhancement in both traffic

scenarios in the downlink case, while regarding the uplink one, the performance of the two schemes is almost identical with a minor decrease in the low and medium traffic cases. It should be noted that the higher gain for the CEPE-enabled scheme is observed in the Web and the Video service types. The observed amelioration in the throughput KPI per Service is due to the enhanced allocation of the users to the respective RATs based on the Services properties and requirements. Note that the Voice/VoIP are practically the same whereas for the more demanding ones we have more benefits.

5.5.2 CEPE 2nd Experiment

Experimentation Scenario and Setup

In this 2nd experiment, once more we use NS-3 in the well-established Shopping Mall scenario. CEPE is juxtaposed against A2A4 RSRQ algorithm. The simulation environment and the overall configuration appear in the following table:

Table 19: CEPE Experiment 2 Simulation details

Simulation Parameter	Parameter Value
Simulation Time	600 sec
Number of UEs	40
Number of LTE Macro cells	2
Number of LTE Femto cells	3
Number of GSM cells	1
Scheduler Algorithm	Proportional Fair
LTE macro & GSM TxPower	35 dBm
LTE femto TxPower	20 dBm
Bandwidth (all cells)	15 Resource Blocks
Path Loss Model	Hybrid Buildings Propagation Loss Model

As already discussed in the presentation of the framework, a network administrator must define and apply specific traffic-related rules, in accordance with the specific network KPIs (rules may change either manually or dynamically according to the status of the network resources and strategy of the operator). In order to assess the validity of the proposed solution, we apply the simplified handover rules as shown in Table 20. In this example, the rules prioritize the targeted cell type during a user's handover taking into account the CEPE defined user profile and associated KPIs.

Table 20: Handover rules applied during evaluation scenario

Rule-Profile #	Service	Mobility	Target RAT
1	Video streaming	Low/Medium	Femto Cell
2	Video streaming	High	Macro Cell
3	Voice call	All	GSM Cell

4	Web browsing	All	Femto Cell
5	FTP downloading	All	Macro Cell
6	VoIP	All	Macro Cell

The main rationale behind these rules is explained below:

- *Rule 1:* Video service demanding in terms of resources (particularly in the downlink). In addition, the Video service is a medium-duration service. As a result, low and medium moving users are unlikely to perform handovers often.
- *Rule 2:* Video users with high mobility are likely to perform multiple handovers if placed in a small cell, thus a macro cell, which has a larger coverage, will serve them in a more efficient way.
- *Rule 3:* The (legacy) GSM cell can handle the voice service due to its limited data rate requirements. This way the LTE cells are offloaded.
- *Rule 4:* Web traffic type is characterized by sparse non-periodic bursts. As a result, frequency of handovers does not influence the performance of such a service type.
- *Rule 5:* FTP is a long-duration service that requires high data rates during the whole file transmission, so we avoid possible handovers –and as a result lost packets or delay overheads- by using macro cells that cover larger areas.
- *Rule 6:* Typically characterized by medium to long duration (and with no very high throughput requirements), it is preferable to place VoIP users in macro cells in order to avoid potential handovers.

In this second series of CEPE experiments, COmpAsS has been deployed as well in order to reinforce the traffic steering process. The UE-side COmpAsS calculates the *Suitability* for the available RATs and base stations. The RAT is thus first selected based on the network side rules presented above, while the specific base station/AP is indicated by the Suitability calculated by COmpAsS.

Results

In the presented scenario, the afore-mentioned traffic types are initially assigned to the UEs of the simulation in a uniform way. Afterwards, we increase or decrease the frequency of specific traffic types, -as it will be indicated in detail in the following paragraphs-, according to the aims of the respective sub-scenario. In order to demonstrate how the system responds to increasing system throughput requirements as well, we have chosen to perform seven sub-scenarios, in which we are gradually increasing the overall required throughput by increasing the percentage of UEs consuming an FTP downloading service. The FTP service has been set-up in our simulations as the most bandwidth-demanding traffic type. The “FTP users” (i.e. the UEs that are consuming only FTP throughput throughout the whole simulation) are increasing from 0 to 70% of the overall users. The KPIs used to demonstrate the performance of our framework is the number of realized handovers, the experienced throughput and the experienced delay. In each sub-scenario different “seeds” are used for the initialization of the user mobility patterns. In detail, the seven sub-scenarios contain [0, 8, 12, 16, 20, 24, 28] FTP users respectively (out of the 40 overall UEs). The FTP service has a downlink rate of 4200 kbps and no considerable uplink rate. In the following analysis the considered KPIs are related to all the scenario users – not only the FTP ones-. As a result, both the

downlink and the uplink metrics are considered. The following figures illustrate the results over the overall number of UEs (mean values) involved in the scenario.

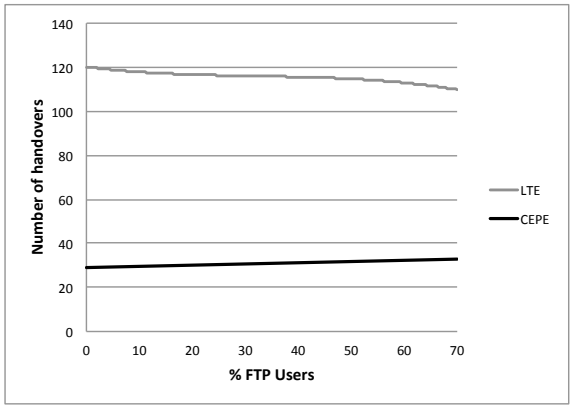


Figure 77: Number of realized handovers (all users)

Figure 77 illustrates the number of handovers performed by all UEs that participate in each scenario. Comparing the results between these two mechanisms show that when applying CEPE mechanisms, UE placement to the respective RATs and cells radically reduces the handovers that take place. A high number of handovers is tightly linked to increased delay; as a result, there is a significant optimization of the delay KPI when we apply the proposed mechanism instead of the normal LTE scheme.

The overall delay that is experienced by the UEs is –on one hand- related to the number of the handovers that was presented in the previous figure, however, additional parameters influence the latency that is introduced. Some of the most important such parameters relate to the UE placement and RAT selection, which -depending on the overall load and the number of associated UEs- may introduce additional delays (e.g. due to the scheduler). Figure 78 illustrates the mean experienced delay in the uplink. CEPE outperforms LTE scheme in almost all sub-scenarios, by 50-150ms.

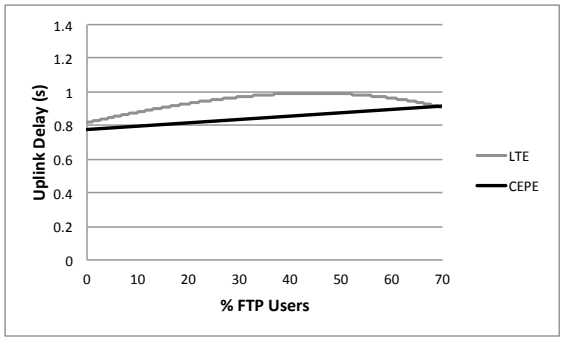


Figure 78: **Mean uplink delay for all users**

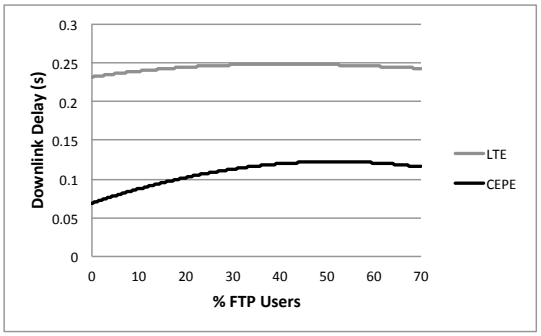


Figure 79: **Mean downlink delay for all users**

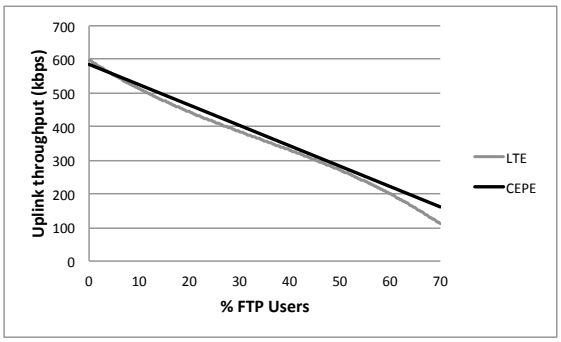


Figure 80: **Mean uplink system throughput**

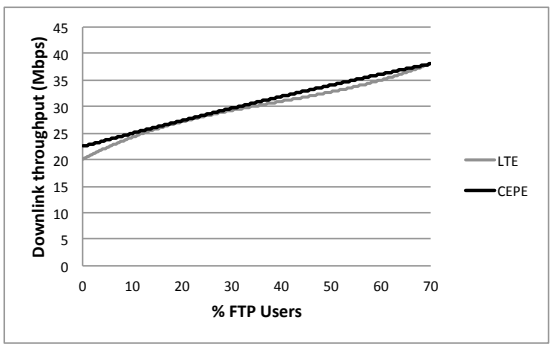


Figure 81: **Mean downlink system throughput**

Although in the downlink (Figure 79) the experienced delay is lower for both schemes when comparing with the uplink, once more the CEPE scheme performs better than the LTE. In all scenarios, the delay that is experienced when deploying CEPE is reduced by more than 50%.

Additionally to the handover frequency and the delay, the throughput metric is also considered as one of the most indicative KPIs for such traffic types.

In the uplink case (Figure 80), although the performance of the two mechanisms is quite similar, it is observed that for higher number of FTP users, the throughput in the CEPE scenario outperforms the LTE case. Similarly, in the downlink case, the throughput is higher up to 3 Mbps in some cases.

6. CIP: A CONTEXT INFORMATION PRE-PROCESSING MECHANISM TOWARDS SIGNALING MINIMIZATION FOR 5G NETWORKS

6.1 Introduction

As already extensively discussed throughout this dissertation, mobility and resource management in 5G mobile communication systems is expected to rely on a huge extent on context-aware dynamic adaptation of resources' provision to heterogeneous user profiles, devices and co-existing Radio Access Technologies. The realization of diverse use-cases involving Machine-to-Machine communications, Internet-of-Things devices and demanding multimedia smartphones will eventually result in a massive amount of contextual data within a complex ecosystem. This ecosystem requires efficient and effective processing on the contextual data by the dedicated network entities in order to extract meaningful network information, while at the same time necessitates refraining from an excessive increase of the signaling overhead due to contextual information exchange. In order to tackle these challenges, in this section, we introduce a novel scheme, the Context Information Pre-processing (CIP) module, an integral part of the previously proposed profiling engine acting as a pre-processing engine aiming to identify and discard redundant or unnecessary data before knowledge extraction. In the context of this paper, we provide a thorough description and analysis of the framework, while in parallel we assess the validity and viability of the solution proposed through extensive experimentation and evaluation scenarios.

6.2 The Content Information Pre-processing Engine: Overview of the solution

The various contextual items collected by a UE can be randomly transmitted to a knowledge extraction entity multiple times as required, thus increasing the signaling overhead induced. In parallel, when dealing with historical information coming from UEs, and/or the core network, we expect that a repeated pattern can be identified due to the strong temporal and spatial nature of their activity. This repeated data pattern –when processed by data analytics schemes– typically does not provide additional knowledge to already available models. Evidently, when considering thousands or millions of coexisting devices this comprises a huge overhead in terms of network and computing resources.

In order to support the communication requirements of 5G networks while in parallel facilitate information exchange among network entities for context extraction, novel, sophisticated solutions have to be developed and deployed. These solutions should address data pre-processing prior to transmission by means of aggregations, data reduction, outlier removal or filtering in order to minimize the number of messages as well as the data size.

CIP acts in an augmenting manner to the previously presented CEPE, i.e., is responsible for minimizing the overall context information that will be forwarded to CEPE for further (pre-)processing. Figure 82 provides a high-level description of the framework. CEPE incorporates both functions of “User Behavioral Profiles Extraction Function”, and “User Behavioral Profiles Storing and Distribution Function”. In this case the CIPs are processing the context information and provide the processed data to the CEPE. The CEPE uses the processed data for extracting the user behavioral profiles, stores the new profiles locally and distributes them to the entities that need them and the CIPs.

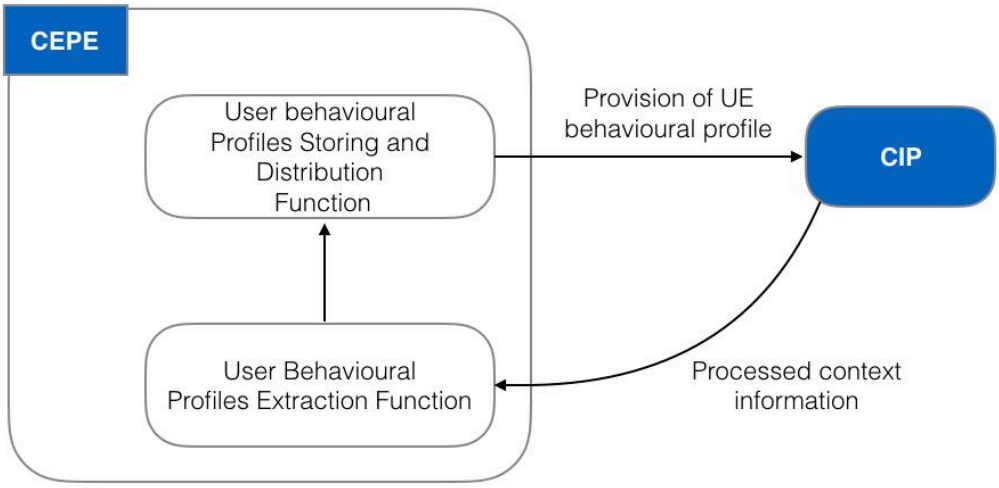


Figure 82: High level description of the CIP-CEPE interworking

The Content Information Preprocessing engine comprises a framework that primarily relies upon data aggregation and pre-processing techniques. The module comprises aggregating and compressing mobile network-related context information per unique identifier, such as the International Mobile Subscriber Identity (IMSI) as well as techniques related to identifying and discarding user profile-redundant or unnecessary context data, before any transmission to CEPE. Except from IMSI, data preparation and compression can be based on other UE parameters, such as location (geographical or network-abstracted area). As discussed earlier, the content is user-related context information (e.g. user traces) aggregated and transmitted for further processing to CEPE. The information that it is collected in the CEPE per user equipment may include, but is not limited to the following categories (detailed information items in Figure 83):

- *Network measurements and information*, comprising but not limiting to received signal strength, RSRP/RSRQ, backhaul link capacity and quality, packet loss, delays, interface information, associated cell ID, Mobile Country Code (MCC), Mobile Network Code (MNC), etc.
- *Mobility Information*, comprising but not limiting to user speed, number of handovers, etc.
- *Service measurements*, comprising but not limiting to accessed service type, accessed service duration, accessed service characteristics (i.e., packet size, packet transmission interval, packet reception interval, uplink and downlink bit rate, acceptable jitter, acceptable packet loss, acceptable packet error rate, etc.)
- *Social information*, comprising but not limiting to age, employment/profession, education, income, gender.
- *User contract information* comprising but not limiting to Contract Id, Signature and Expiration dates,
- *Charging Information*, charging data records, comprising but not limiting to charging model, available credits,
- *UE description information*, comprising but not limiting to available battery, maximum battery charging, device central processing unit description, memory, operating system, screen size, screen resolution, power/energy information (e.g., battery consumption rate, battery level, current CPU, current memory, information about protocols supported by the user equipment (e.g., Type of protocol, Required memory, Required CPU), data describing the physical interfaces offered by a device (e.g., uplink rate, downlink rate, round trip delay, errors, packets sent, packets received, etc.), etc.

- Timing information.

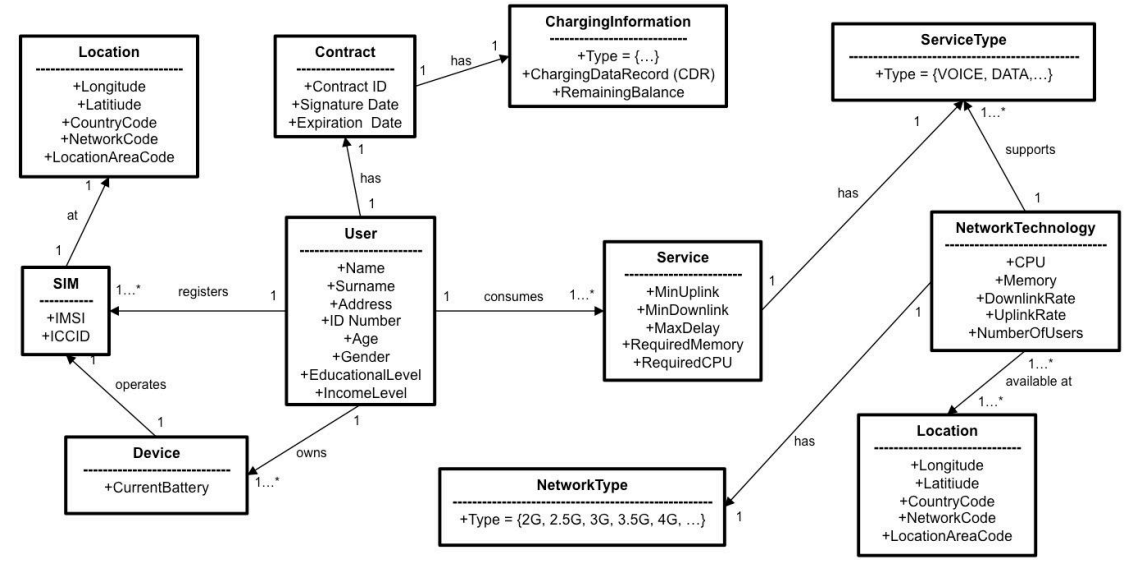


Figure 83: CIP detailed Information model

Figure 84 shows an exemplary implementation of the proposed scheme in a LTE/LTE-A network for gathering the information to the CEPE, which as mentioned afore is a logically centralized network entity and communicates logically with the CIPs, either directly, or through other CIPs. As depicted CIPs may reside in all the entities where the afore-mentioned information elements could reside. In other potential implementations, the CIPs could reside to any networking device that could provide information related to user context, including but not limiting to servers, databases, access nodes, UEs, etc.

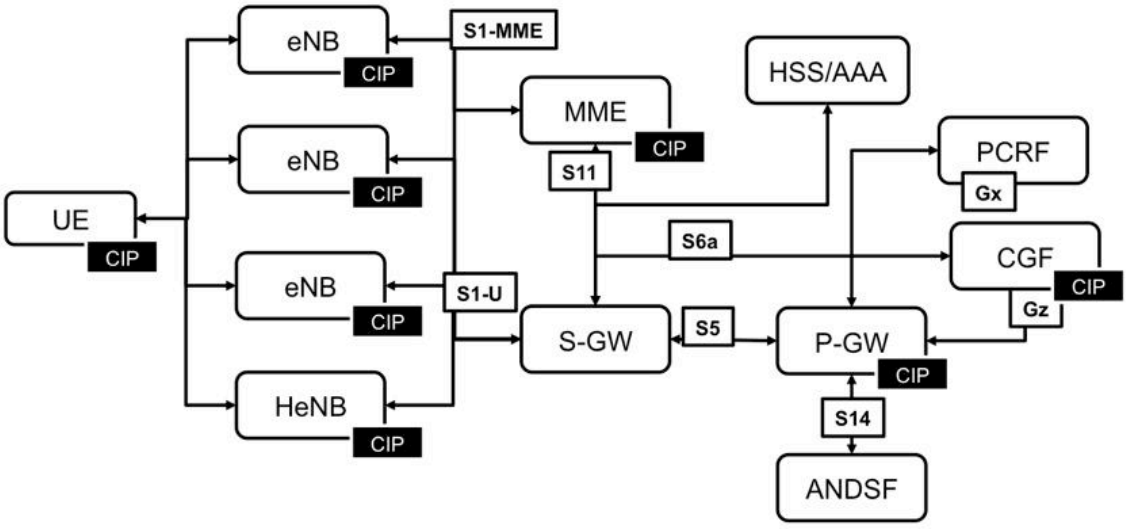


Figure 84: CIP potential distributed deployment in LTE - EPC

Each CIP aggregates context information per a unique user identifier and in the aggregated information it identifies redundant or unnecessary data and discards it. The aforementioned unique identifier may be the user IMSI or any other unique identifier for the user, so as to indicate that this context information refers to a certain user.

Once the CEPE has generated the user behavioral profiles, it distributes them to the networking elements that may contain or obtain information related to a certain user. The distribution of the profiles to the networking elements may be on demand, or automatically (e.g., periodically, or when the CEPE has updated user behavioral profiles, etc.) from the

CEPE. As described afore, the user predicted behavior includes the preferences/predictions in relation to several communication characteristics, including but not limiting to expected services to be accessed, the access rate, access duration, user mobility, etc. The user behavior prediction may be based only on the user behavioral profile, or on one of the user behavioral profile and real-time information. In one potential implementation, where the user behavioral profile relates with real time information the real-time inputs (e.g., time/time period, location, charging, battery level, etc.) could be combined for making more accurate prediction regarding the user behavior. Table 21 presents one exemplary implementation of the user behavioral profile that predicts the accessed service and the user mobility, which combines real time information for a certain user, under certain preconditions (i.e., location, weekday, time/time period, battery status, charging status). The user behavioral profile that is used for making predictions regarding the user behavior under the preconditions will called from this point active user behavioral profile or simply active behavioral profile.

Table 21: Two-dimensional exemplification of the behavioral profiles

Location	Day	Time	Battery Status	Charging Status	Predicted Service to be accessed	Predicted Mobility
Stadium	Mon-Fri	9:00-18:00	Fully Charged	No Credits remaining	Short Voice Calls	Low Mobility
Stadium	Mon-Fri	9:00-18:00	Fully Charged	Pre-paid credits remaining	Long Voice Calls	Low Mobility
Shopping Mall	Mon-Fri	9:00-18:00	Fully Charged	No Pre-paid credits remaining	Only Web	High Mobility
Shopping Mall	Mon-Fri	9:00-18:00	Fully Charged	Pre-paid credits remaining	Only Web	Medium Mobility

Then, the corresponding CIPs that reside in these networking elements, will consider the active behavioral profile and the respective predicted behavior and will discard any redundant information. By the term redundant information, we refer to any information that is in accordance to the active behavioral profile. Figure 85 provides an exemplary implementation of the distribution of the profiles in LTE/LTE-A networking elements. However, this implementation could be applied to any type of networking deployments. The distribution of the profiles could be either via direct communication or through other networking elements and the respective CIPs. In the depicted exemplary implementation (Figure 85) the behavioral profiles are distributed through one “primary” CIP that resides in a certain networking element (in this exemplary implementation in the Mobility Management Entity - MME).

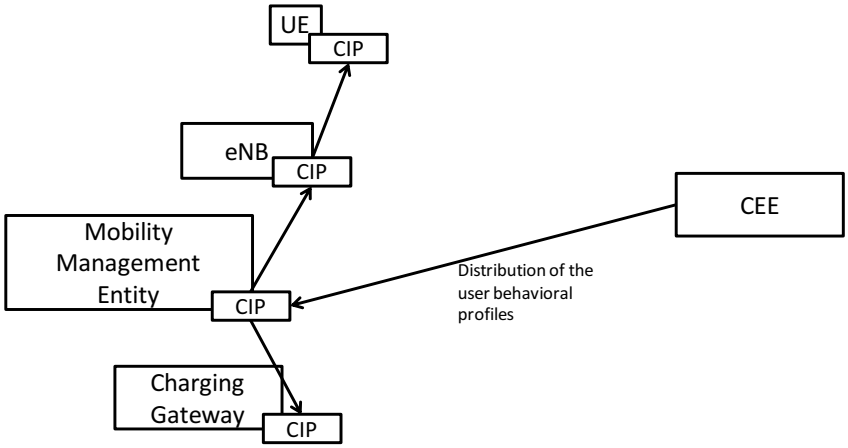


Figure 85: Potential implementation of active user behavioral profiles’ distribution

Figure 86 details the CIP process, when deployed in a network entity acting as contextual information source:

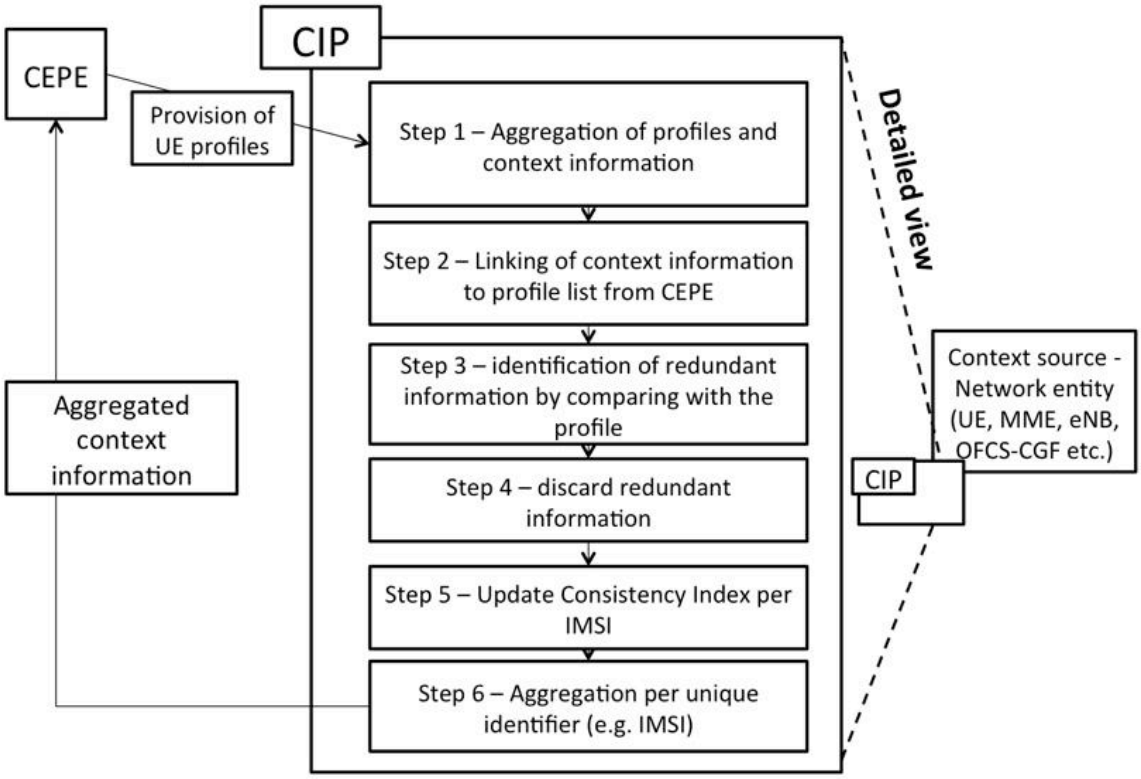


Figure 86: CIP operation

1. *Aggregation of profiles and context information:* via the respective interfaces, context information is pulled/ pushed to CIP.
2. *Linking of context information with the profiles provided by CEPE:* for each connected IMSI, CIP makes an association with the respective profile that has already been acquired by CEPE and stored in CIP.
3. *Context information comparison with the respective profile - identification of redundant information:* if the information item that is evaluated is consistent with the respective profile, the item is identified as redundant
4. *Redundant information removal:* Redundant information is discarded since it offers no additional insight to the CEPE data-mining phase.
5. *Consistency Index update:* every time an information item is identified as redundant, the CI is updated accordingly
6. *Aggregation per unique identifier (e.g. IMSI)*: the filtered context information items are aggregated per identifier in order to be transmitted.

Since each user may have more than one behavioral profiles related to the context information (e.g., location, time periods, date, battery level, etc.), and each CIP may have several user behavioral profiles that refer to one or more users, the CIP when it has to provide context information to the CEPE (Step 1 – Figure 86), then it links the context information with the respective user behavioral profile (Step 2 – Figure 86) for the identification of the active behavioral profile. This implies that the CIP will map the context

information to be transmitted with the predicted behavior of the user under certain preconditions that refer to the process that produced the context information. Then, by comparing the context information of the user, which is the actual behavior of the user with his predicted one (represented by the user behavioral profile), the CIP identifies the redundant information (Step 3 - Figure 86). Specifically, when a user complies with his predicted behavior, this information is considered redundant and may be discarded (Step 4 - Figure 86). When the user does not comply with his predicted behavior, then this information is not redundant and has to be collected by the CEPE so as to be considered in the user behavioral profile extraction process from the "User Behavioral Profiles Extraction Function". Each time redundant data is discarded, a consistency index (CI) is updated for the associated user identifier. Only not redundant data and the Consistency Index are transmitted to the CEPE either directly or through other CIPs (Step 5 & 6 – Figure 86).

In one exemplary implementation, when the user behavior is described by the user mobility and the service that the user accesses in a certain time period, then his behavioral profile could be as depicted in Figure 87 for two certain time periods. In this exemplary implementation, other parameters could be considered apart from the time periods for making more accurate predictions such as the location, charging, battery level, etc. Then, when in time period A (e.g., 9:00-12:00) of a Day A (e.g., Monday) she is predicted to have Mobility profile A and access Services described by profile A (e.g., long voice calls). The UE behavior may move towards three directions:

- she will follow his behavioral profile,
- she will have a totally different behavior, or,
- she will be inactive.

IMSI A	Mobility A	Service A	TimePeriod A	Day A
IMSI A	Mobility A	Service B	TimePeriod B	Day A

Figure 87: User behavioral profile for two time periods

In this exemplary implementation for enabling the CEPE to build the behavioral profiles using context extraction mechanisms, if CIP are not applied, all the user actions should be recorded in the user side with the certain timestamps and be transmitted to the CEPE.

Figure 88 describes the records in a UE where the behavior of a UE with IMSI A, is described in terms of accessed services and mobility for several timestamps (*TS1-TS12*). Then, the CIP will identify the information that is redundant and could be discarded. When the UE follows his behavioral profile, then he does not need to transmit this information, since he complies to his predicted behavior. Instead she will increase the CI, which is a counter capturing the times that the UE is compliant to his predicted behavior. This is necessary for the CEPE to be able to reconstruct the overall user behavior (including the inactivity and the deviations from the profile). Specifically, in the presented example, the UE CIP would transmit a structure which would contain, (a) only once the UE IMSI (IMSI A), (b) the Consistency Index for the times that the UE followed the profile (i.e., 5), and (c) the 5 times when it deviated from the profile (Figure 89).

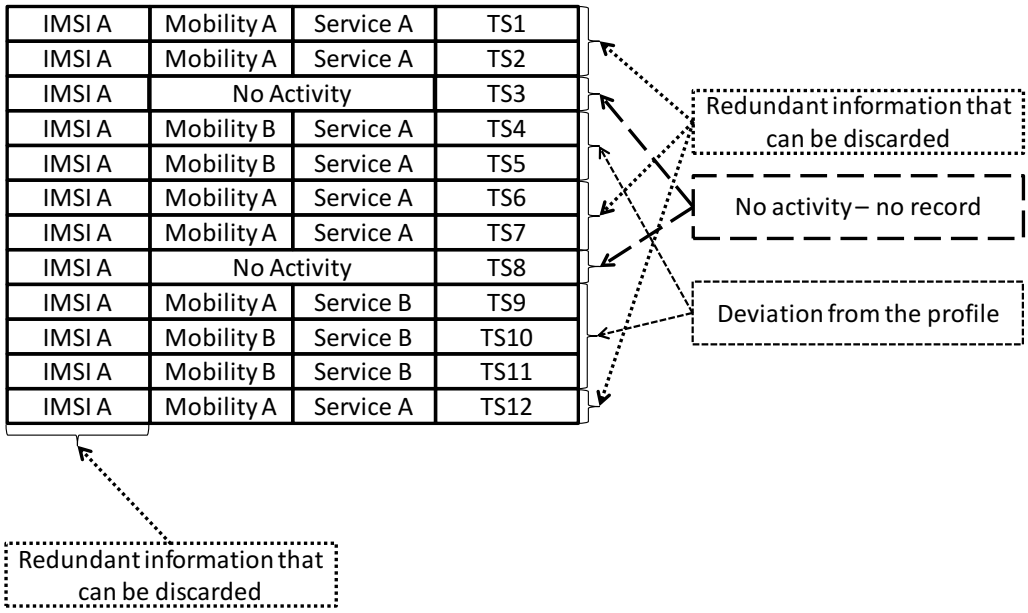


Figure 88: Information fields description

IMSI A			CI
Mobility B	Service A	TS4	
Mobility B	Service A	TS5	
Mobility A	Service B	TS9	
Mobility B	Service B	TS10	
Mobility B	Service B	TS11	

Figure 89: Transmitted information for the UE CIP using the CI

This information is being transferred the other CIPs (e.g., the CIP of the eNB) either periodically or on demand. The other CIPs may collect several records; they will aggregate this information, and they will discard redundant information before sending it to other CIPs.

In one alternative implementation, the CEPE may provide to the CIPs the predicted behavior of the user (Figure 85) and one table with other predicted behavioral profiles, as shown in Figure 90.

Mobility A	Service A	Profile ID 1
Mobility B	Service A	Profile ID 2
Mobility C	Service A	Profile ID 3
Mobility A	Service B	Profile ID 4
Mobility B	Service B	Profile ID 5
Mobility C	Service B	Profile ID 6
Mobility A	Service C	Profile ID 7
Mobility B	Service C	Profile ID 8

Figure 90: Predicted behavioral profiles that CEPE provides to the CIP

In this implementation when the UE follows his behavioral profile the CI will be transmitted as in the previous implementation, but when the user deviates from his behavioral profile instead of sending the behavior, only the Profile ID that captures the behavior will be provided, as shown in Figure 91.

IMSI A		CI
Profile ID 2	TS4	
Profile ID 2	TS5	
Profile ID 4	TS9	
Profile ID 5	TS10	
Profile ID 5	TS11	

Figure 91: Transmitted information from the UE CIP using the CI and the Profile IDs provided by the CEPE

Other alternative implementations could include the transmission of specific counters for every behavioral profile instead of separate, as shown in Figure 92.

IMSI A	CI
Profile ID 2	Counter for Profile ID 2
Profile ID 4	Counter for Profile ID 4
Profile ID 5	Counter for Profile ID 5

Figure 92: Transmitted information using counters for the Profile IDs when the UE deviates from his profile

Since the information does not have to be provided to the CEPE in real time, but it may be provided offline, this enables the aggregation of information on a per CIPs basis as well. Specifically, the context information may be transferred to the CEPE via several CIPs. This facilitates the above procedure to take place in all the CIPs until the context information reaches the CIP. Each CIP provides the aggregated information to another CIP so as to be forwarded to the CEPE is implementation specific. Potential implementations include but are not limited to periodically transmission of aggregated information, transmission after certain amount of information has been aggregated, etc. This enables each CIP to aggregate information on a per unique identifier basis. For example, if a CIP residing in an eNB aggregates context information for a UE that is associated to this eNB; then the associated cell ID could be discarded, since it is redundant. This process is redundancy identification per CIP identifier. Similar redundancy identification apart from per networking element identifier may be extracted per location (e.g., Mobile Country Code (MCC), Mobile Network Code (MNC), date, time, etc. One exemplary implementation of this procedure is depicted in Figure 93 for an LTE/LTE-A network. As it is depicted, context information (upon removing redundant information) is being transferred from the CIP that resides in the UE to the CIP that resides in the eNB. The CIP that resides in the eNB aggregates context information and removes redundant information on per eNB identifier, per location and per time basis. Then (after a certain time interval or on a per aggregated data volume basis) the aggregated information (that contains context for many UEs) is being transferred from the CIP that resides in the eNB to the CIP that resides in the MME where the same procedure is followed and redundant information is being discarded. As it is shown in the figure, the MME may receive information from CIPs that reside in other types of networking elements (such as charging severs in Figure 93). Then the overall aggregated information for many CIPs is being transferred to the CEPE.

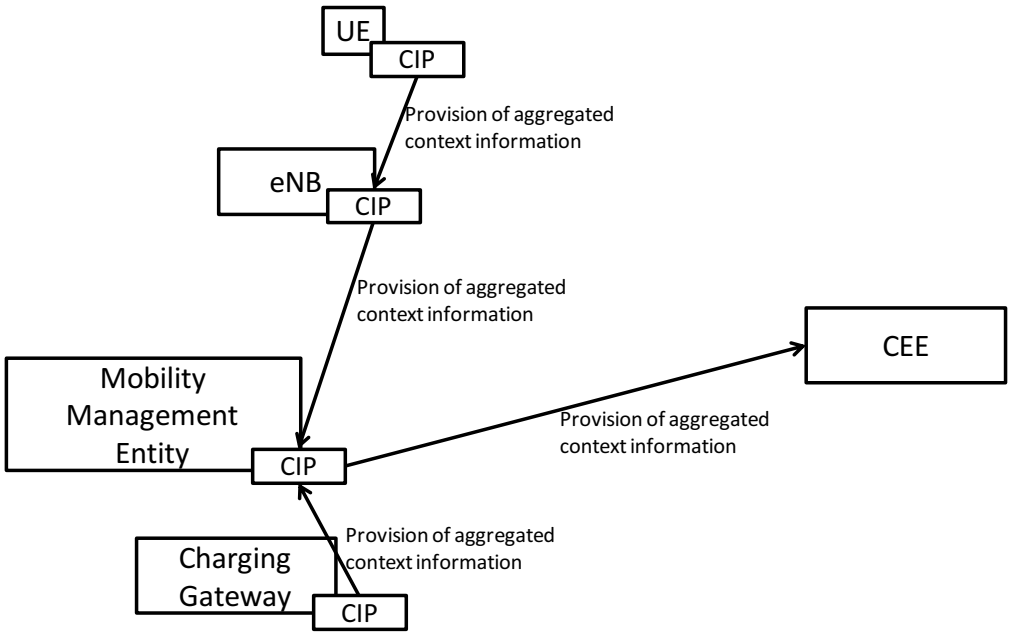


Figure 93: Process of redundancy removal on a per CIP identifier level

6.3 Analytical CIP evaluation and overall system overhead quantification

In this section, we provide the evaluation of CIP using an analytical approach. For the sake of homogeneity, the use case, which is applied, is once more the Shopping Mall scenario, as this was presented in all previous evaluation sections of COmpAsS and CEPE mechanisms.

We assume that the operator is performing profile-based data analytics using CEPE. Various parameters related to context information should be aggregated from numerous network entities, each of which is characterized by a different payload as well. For this reason, we provide Table 22 that shows the payload that we used in our evaluation schema; for each context item, we also provide a reference, in which detailed insights are presented.

Table 22: Payload per context information parameter

Parameter	Payload	Source Network Entity	Reference
IMSI	64 bits	eNB, MME	[205]
Cell ID (+MCC+ MNC)	32 bits	eNB, MME	[206]
Timestamp (TS)	32 bits	All	[207]
Mobility (Mobility State Estimation based on Handover counters)	8 bits	MME	[208]
User Charging Data Record (CDR)	64 KB to 100 MB	OFCS (Offline Charging System) – CGF	[209]
Contract type (in SPR – Subscription Profile Repository)	4 bits	PCRF – PGW	[210]

Charging status (enough credit/no credit)	1 bit	Online Charging System (OCS) – PGW	[211]
UE Battery Status (normal/low)	2 bits	UE	[212]

Assuming that, under multi-hop operation, individual resources are required for each individual link/hop, in a distributed CIP deployment, the signaling overhead estimation will eventually depend on the network entities, where CIP and CEPE are deployed, corresponding to different number of network hops and direct interfaces between the source and destination of the context information items.

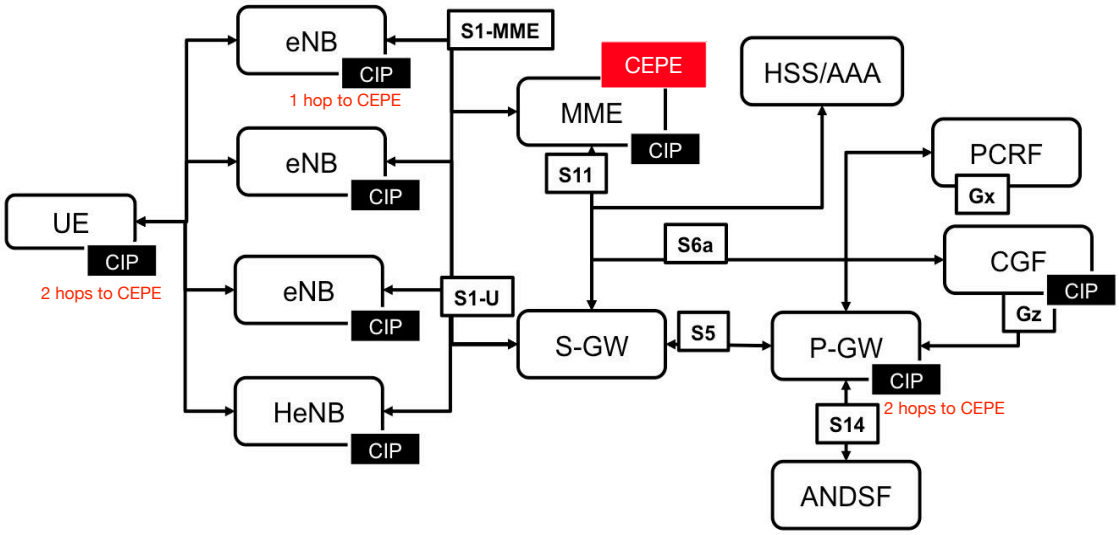


Figure 94: CEPE and CIP deployment in the network during the evaluation

Figure 94 shows an example, in which CEPE is deployed on the MME; CIP_{UE} is two hops away from CEPE, CIP_{eNB} is 1 hop away, while $CIP_{P\text{-}GW}$ is also two hops away. A similar scenario may of course be applied in the same way for N hops between the several network entities, in case CEPE and CIP follow a different deployment.

In our evaluation scenario towards the proof of the CIP concept, we assume –for the sake of simplicity of calculations- that all CIP – CEPE communication links are implemented with direct interfaces. We compare the proposed approach against two mechanisms by quantifying the overhead induced to the basic daily signaling due to their application (Table 23).

Table 23: Juxtaposed schemes

Proposed scheme	Pre-processing	Redundant information identification/discard
Standard	None – all information is aggregated to be transmitted	None

Semi-optimized	Basic pre-filtering (pre-aggregation per IMSI and location)	None

The format of the transmitted context information comprises all the required information items, as discussed earlier and is described as follows:

{ (IMSI, location, TS), MSE / EPS Bearer ID / contract type / charging status / battery status }

The IMSI-location-TS triplet comprises a key and is always transmitted along with one or more of the remaining context parameters (e.g. MSE, EPS bearer ID, etc.). We define Equation 2 in order to calculate the overall signaling cost imposed by the transmission of the afore-described context information parameters.

$$O = \sum_{i=1}^{N}\sum_{j=1}^{M}[(1-P_c)x[S(I_i)\,x\,Fr(I_i) + S_{CDR}] + S(I_{static}) + S_{cc}]$$

Equation 2: Signaling overhead calculation – for 1 hop distance between CIPs - CEPE

where *O* is the overall signaling overhead*, N* the number of users connected to the system, M the number of profiles/behaviors that the UE exhibits, P_c the consistency percentage that describes the portion of the acquired context information, which is consistent with the existing user profile*,* $S(I_i)$ the payload of each one of the *i* context parameter {TS / MSE / contract type / charging status / UE battery status}*,* $Fr(I_i)$ the transmission frequency of each one of the *i* parameter {TS / MSE / CDR / contract type / charging status}*,* $S(I_{static})$ the payload of the static tuple {IMSI, location} (64 + 32 bits = 96 bits)*,* S_{CDR} the message of the varying CDR file sizes (1/10/100 MB) and *Scc* the size of the Consistency Counter.

The evaluation that follows illustrates the system overhead imposed during the context data acquisition and for varying scenarios of UE profile consistency. We assume that the consistency percentage - P_c - varies from 10 up to 90%, as shown in the first chart (Figure 95).

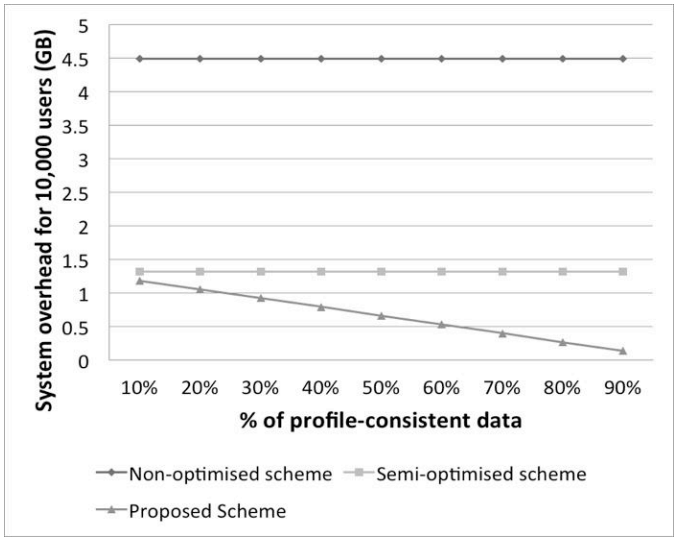


Figure 95: System overhead with varying profile consistency

The second evaluation chart (Figure 96) illustrates the gain of the proposed scheme – CIP- against the two afore-mentioned standard mechanisms. As it can be seen on the X-

axis of Figure 96, we attempt to assess the CIP's gain for different cases, in terms of the CDR file sizes, which often may vary; in our evaluation we use varying sizes, from very small to considerably large: 1, 10 or 100 MB according to respective references ([209]). Later, we show, how the varying CDR size also influences the evaluation results.

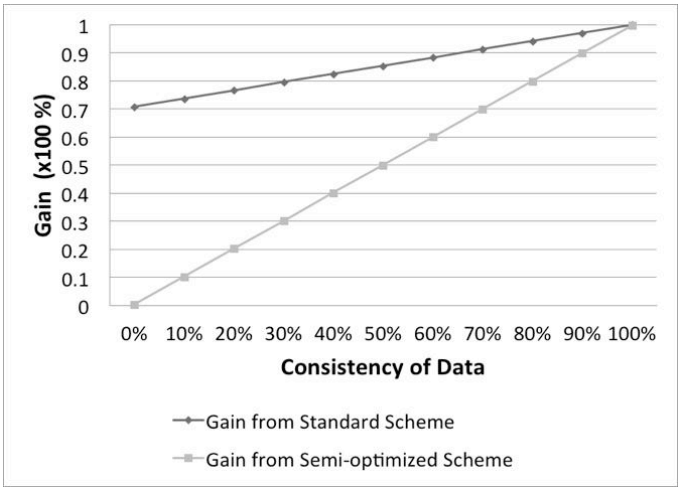


Figure 96: CIP's gain over the two selected schemes

For the evaluation, we consider some of the most challenging 5G use cases (Figure 97), which follow the METIS project specifications ([182]) and correspond to varying number of UEs in the network environment; traffic jam, shopping mall, stadium, open-air festival are some indicative examples. The results of the initial 3-fold evaluation that was described above are illustrated in the following charts.

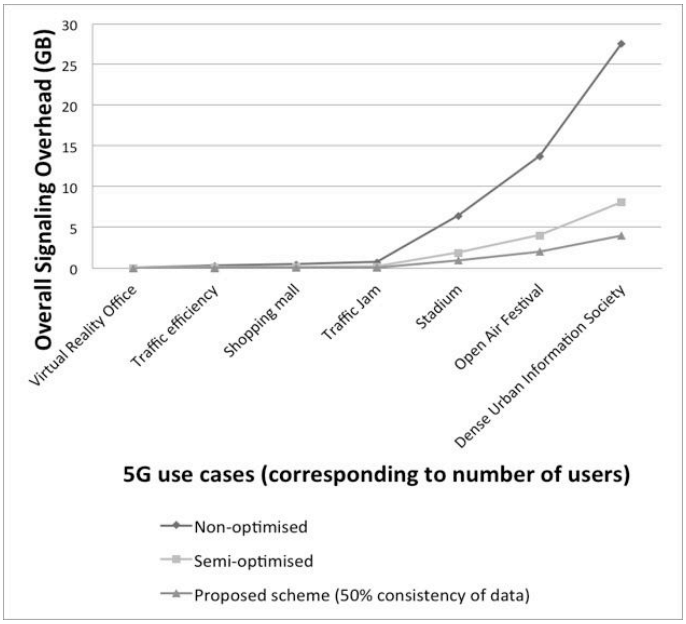


Figure 97: Signaling overhead for different 5G use cases

As depicted in the graphs, the gains of applying CIP during the context information acquisition and transmission between the different network entities may vary from small to huge. The most important parameter that influences the extent of the signaling overhead minimization is the data consistency in terms of the already available user profiles.

In other words, as one would expect, the more consistent the context data is with the CEPE-produced user profiles, the more the overhead approaches zero when CIP is applied (Figure 95, Figure 96). In all cases, the semi-optimized scheme performs better

already than the standard scheme, which transmits all the information towards the Knowledge Extraction engine (CEPE), while both of these schemes our outperformed by CIP.

In the last part of the evaluation of CIP, we illustrate how the overall overhead is influenced by the varying CDR daily payload, which –as described earlier- may vary between 1 to 100 MB. The graphs (Figure 98 - Figure 99), which follow, illustrate the results for 3 distinct CDR sizes, i.e. 1, 100, and 100 MB, and 2 distinct profile consistency (P_c) values, i.e. 90% (very consistent) and 10% (barely consistent).

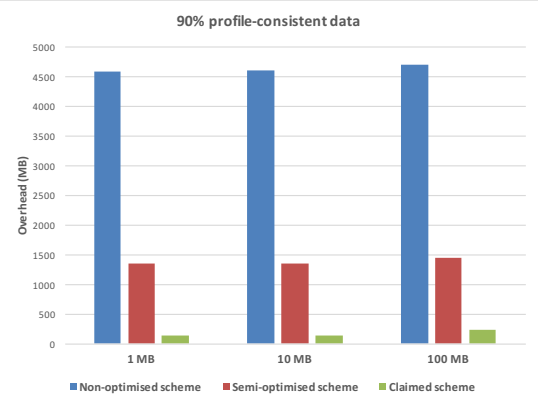


Figure 98: Overhead for varying CDR sizes and 90% data consistency

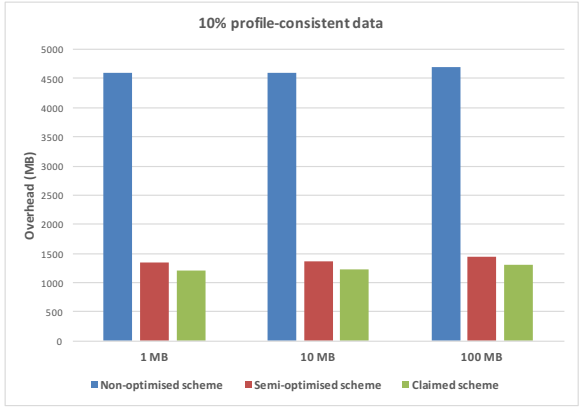


Figure 99: Overhead for varying CDR sizes and 10% data consistency

As it is expected, the data consistency is of utmost importance, even for large CDR sizes. In case of 90% profile consistent data the difference between CIP and non- optimized or semi-optimized schemes is huge. On the other hand, when the profile data is hardly consistent – which translates to the need for transmission of almost all data acquired-, CIP –although still slightly better- performs similarly to the semi-optimized scheme.

Towards 5G systems, one of the greatest challenges is the enormous increase of the number of devices to be handled by the network in an efficient way. Figure 97 illustrates the great gain that results from the application of CIP, particularly in use cases, such as an Open Air Festival, or a Dense Urban Information Society, where the numbers of users begin to increase almost exponentially.

7. CEPE & COmpAsS interworking in 5G architecture

7.1 Introduction

The main part of this thesis has been carried out based on the System Architecture Evolution (SAE) and Evolved Packet Core (EPC) for the 4G Long Term Evolution system. In the meanwhile, several EPC entities have evolved towards the 5th generation of the system, as described in the respective 3GPP working items ([154]), as well as other standardization bodies. This section carries out a detailed discussion on how the proposed framework will operate in the 5G system architecture, based on 3GPP's latest standardization work. Several new network entities, functions and features are being introduced towards 5G and in this discussion, we showcase the validity of the proposed work and the feasibility of a potential integration in the 5G System, according to the very latest work done by 3GPP. In parallel, a discussion on the CEPE – COmpAsS interworking takes place, attempting to showcase the complementary nature of the network-oriented CEPE with the purely UE-oriented COmpAsS and how the assets and shortcomings of each one can provide a holistic context-based traffic steering framework.

7.2 CEPE and COmpAsS interworking

As already mentioned earlier, the core network-based CEPE with the UE-controlled COmpAsS are able to operate independently –and with a different approach- with regard to the RAT selection and traffic steering policies. COmpAsS operates on the UE side, and -based on a multi-criteria fuzzy logic based scheme- calculates the suitability of the available RATs and cells/APs for each one of its active sessions (applications/services/etc.) in almost real-time. On the other hand, CEPE operates as standalone Core Network function/entity, which builds upon historical context information and patterns in relation to the UEs, the network, the consumed applications and services, etc. and generates User Profiles, on which, prediction for future situations is based.

However, the two mechanisms target the same objective (applying optimized radio resource management and traffic steering) from divergent perspectives, which however, seem to be complementary and can be combined –since they are deployed in different network components-. The complementarity of the two schemes is federated by the fact, that -as happens in the majority of context-based mechanisms-, the benefits have always to confront numerous drawbacks and disadvantages.

COmpAsS operation provides knowledge of the current situation of the network status – in the vicinity of the UE- and, based on the respective service requirements, it triggers a session setup (or a handover in case of an on-going session) to the most appropriate access technology and specific cell/AP; besides, COmpAsS minimizes unnecessary execution of control functions based on the mobility of the user, the requirements of a particular service/application, as well as the load of the network.

Nevertheless, certain drawbacks refrain COmpAsS from being capable of supporting the overall traffic steering and RAT selection operation in a holistic manner: **The UE-assisted nature of the specific scheme, results in a solution, which on the one hand optimizes the resources provided for specific UEs, however, on the other hand, lacks any capability of an overall planning or overview of the status of the network, leading requests from the UE side impossible to address in very demanding situations, such as dense deployments.** Novel traffic engineering approaches in forthcoming, challenging 5G environments will require solutions, which will act also partially from the UE side for an efficient, real-time network probing, however, the final network-side decision making will be of utmost importance. Furthermore, COmpAsS acts in a reactive manner; the *context* acquired and processed by COmpAsS refers to recent real-time behavior of the UE; this translates to **slow convergence** in cases of diverse UE

behaviors in small time frames, consecutive calculations from the UE side every time the context is modified (such as mobility changes), etc. Last but not least, a *context*-based scheme on the UE side –no matter its efficiency- implies **additional signaling information, -specifically for ultra-dense environments, where hundreds/thousands of UEs coexist among numerous base stations/Aps -**, which makes the solution inefficient in terms of energy consumption – one of the most crucial aspects of UE-based solutions-, **even if several energy-related optimizations** have taken place in COmpAsS design and implementation.

Contrary to COmpAsS reactive nature, CEPE, on the other side, is a proactive Core Network based entity/function, which aggregates context information related both to the UEs (device characteristics, behavior profile, app usage, etc.), as well as the network status, and builds prediction models and user profiles, based on patterns, which are identified in this aggregated context over a certain amount of time. The network traffic demands prediction is directly linked to one of the strongest CEPE advantages: via the holistic picture of the network over long-time frames: this provides the network administrator a framework of utmost importance, facilitating traffic engineering operations, load balancing, etc.

As already highlighted, CEPE's longer term context processing differs considerably from COmpAsS's, which applies real-time monitoring and decision-making. This one the hand, provides the ability to predict user future behavior, in terms of both mobility, as well as service usage; on the other hand, however, this operation implies **certain weaknesses** for specific/examples: **a new UE connects to the network, which CEPE is not trained for,** or there is a **deviation from the existing profile** (a UE enters a new area, in which CEPE is still not trained or an event occurs (e.g. new type of application/service launched), which deviates considerably from the profile that CEPE has built for the particular UE, etc.). Another shortcoming identified relates to the **limited ability of CEPE operation to perform micro monitoring in real-time** and identify the specific cell/AP, which should be selected for a specific UE –in case of multiple options-. This information is however acquired by COmpAsS, which scans when required the UE's environment in real-time; thus, the two schemes may as well combine context information towards the optimal selection.

To sum up, certain advantages and drawbacks of the operation of the two, primary context-based schemes (i.e. CEPE and COmpAsS) prove the validity and optimization that a solution, which combines their operation in a parallel and coordinated manner, would offer.

- In the CEPE-COmpAsS combined operation, CEPE provides the main resource allocation policies, slice and access traffic mapping per UE profile and distributes them among the relevant network entities, which are responsible for forwarding those policies to the network and the UEs.
- Whenever, an event occurs, which deviates from CEPE's profiles and policies or there is no profile information for a UE (new application type previously not profiled, new femto cell deployment, new UE with no mapped profile, etc.), the network receives COmpAsS's information for the specific, available access technologies, based on real-time context information.
- Moreover, COmpAsS is able to fine-tune CEPE's policies regarding specific APs/femto cells, etc., as CEPE primarily provides high-level policies, but does not link them to specific available cells/APs.
- Additionally, COmpAsS's suitability assessment is forwarded back to CEPE as additional context information for the feedback-based policies refinement.

- Last but not least, CEPE can be exploited to fine-tune the load information request rate of COmpAsS to ANDSF, and/or other relevant context-aware components (if the rate of requests is low, the network will take care of everything through load balancing but no optimal performance will be achieved since the UE may ask for BSs that are loaded even if the RSRQ from a new BS is a better one; on the other hand, if the rate is very high, the UE performs an optimum filtering of unneeded actions, however there is a high increase in wasted resources through many connections to receive control information).

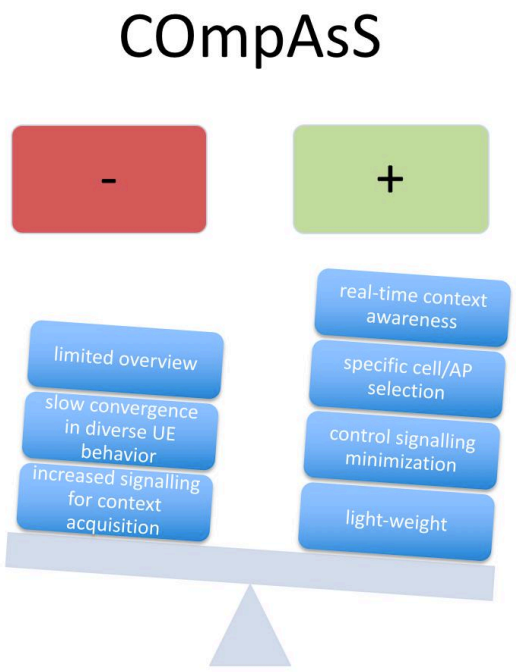


Figure 100: COmpAsS shortcomings and advantages

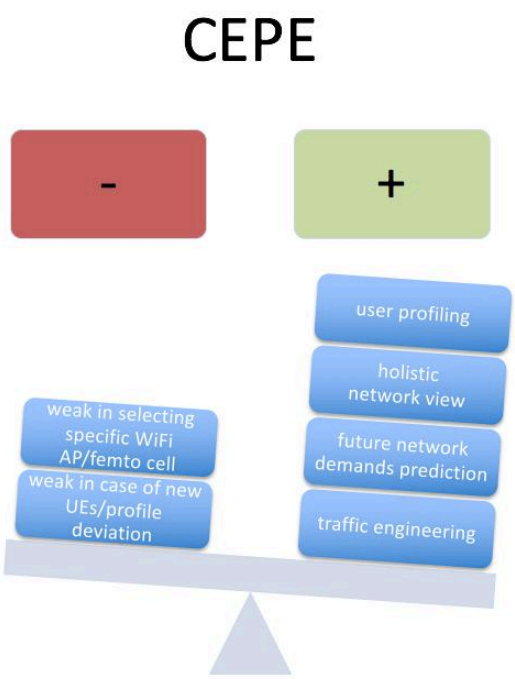


Figure 101: CEPE shortcomings and advantages

7.3 The proposed framework in the 5G System architecture

7.3.1 3GPP's NWDAF: Network Data Analytics Function, NSSF: Network Slice Selection Function and RCAF: RAN Congestion Awareness Function

Continuing the previous discussion about the CEPE-COmpAsS interworking, in this section, we provide a detailed discussion on the proposed framework's compatibility with the 5G network architecture and its core network entities, as well as potential shortcomings and required extensions, which will be required in order to support the required functionality, both for the Core Network's part (CEPE), as well as the UE-based part (COmpAsS).

In [154], 3GPP introduces the 5G network core architecture. Some key principles and concepts target to:

- Modularize the function design, e.g. to enable flexible and efficient network slicing.
- Enable each Network Function to interact with other NF directly if required. The architecture does not preclude the use of an intermediate function to help route Control Plane messages (e.g. like a DRA).
- Minimize dependencies between the Access Network (AN) and the Core Network (CN). The architecture is defined with a converged core network with a common AN

- CN interface which integrates different Access Types e.g. 3GPP access and non-3GPP access.

- Support "stateless" NFs, where the "compute" resource is decoupled from the "storage" resource.
- Support concurrent access to local and centralized services. To support low latency services and access to local data networks, UP functions can be deployed close to the Access Network.

The reference 5G architecture is illustrated in the following figure (Figure 102):

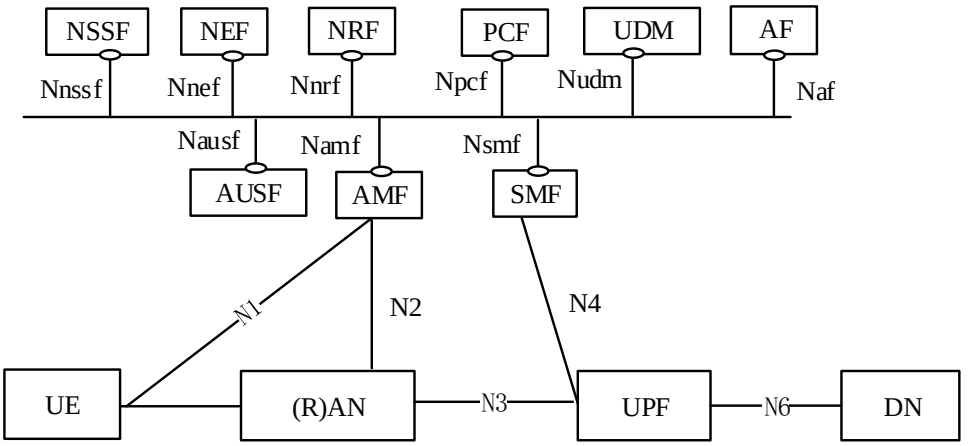


Figure 102: 5G System Architecture

One of the novel core network functions, which are introduced in [154] is a dedicated data analytics function (**NWDAF**). According to the latest standardization discussion, the NWDAF will be providing load level information on a network slice level. NWDAF provides slice specific network data analytics to the Policy Control Function (**PCF**), as well as the Network Slice Selection Function (**NSSF**) over their newly specified interfaces (i.e., *Nnwdaf*, *Nnssf* and *Npcf*). PCF will be using the NWDAF input for optimizing the policies assigned to each UE and its respective on-going sessions and data flows, while **NSSF will utilize the NWDAF's input in order to maintain the optimal UE – slice mapping for the diverse types of UEs and data flow characteristics**, which will be coexisting in certain areas. Besides, PCF and NSSF, also TSSF (Traffic Steering Support Function) will essentially receive NWDAF's output for optimized traffic steering.

CEPE, as a network data analytics and prediction engine is essentially a module, capable of supporting fully this functionality by serving as NWDAF instance in the forthcoming 5G architecture. Apparently, different data mining engines may be deployed by different operators, focusing on specific context information items or particular verticals/business scenarios. The diverse information items, aggregated and processed by CEPE have been analyzed in the respective previous section. Mapping this information, with the afore-presented 5G architecture, several 5G network components, which generate valuable context information, will forward it to CEPE/NWDAF: Unified Data Repository (UDR), Access Network Discovery and Selection Function (ANDSF), PCF, RAN Congestion Awareness Function (RCAF), etc.

RCAF is also a vital component for COmpAsS context information acquisition related to one of its five core context parameters, i.e. the real-time AP/cell load information. COmpAsS will be receiving information related to the potential congestions related to specific APs/cells –besides the ANDSF, which has already been described in previous section- and will calculate the respective *Suitability* of each AP/cell.

7.3.2 3GPP's ATSSS: Access Traffic Steering Switching and Splitting

Another latest 3GPP's working document ([155]) defines a new network function, namely the Access Traffic Steering Switching and Splitting (ATSSS), which is responsible for the management of the different UE flows over the available access technologies. The three main operations, supported by the ATSSS are –as also mentioned in its name-:

- **Access Traffic Steering:** The procedure that selects an access network for a new data flow and transfers the traffic of this data flow over the selected access network. Access traffic steering is applicable between 3GPP and non-3GPP accesses.
- **Access Traffic Switching:** The procedure that moves all traffic of an ongoing data flow from one access network to another access network in a way that maintains the continuity of the data flow. Access traffic switching is applicable between 3GPP and non-3GPP accesses.
- **Access Traffic Splitting:** The procedure that splits the traffic of a data flow across multiple access networks. When traffic splitting is applied to a data flow, some traffic of the data flow is transferred via one access and some other traffic of the same data flow is transferred via another access. Access traffic splitting is applicable between 3GPP and non- 3GPP accesses.

The ATSSS is distributed over different network entities, such as the UE, the UDR, the SMF and the PCF. The ATSSS architecture contains the following functional elements (Figure 103):

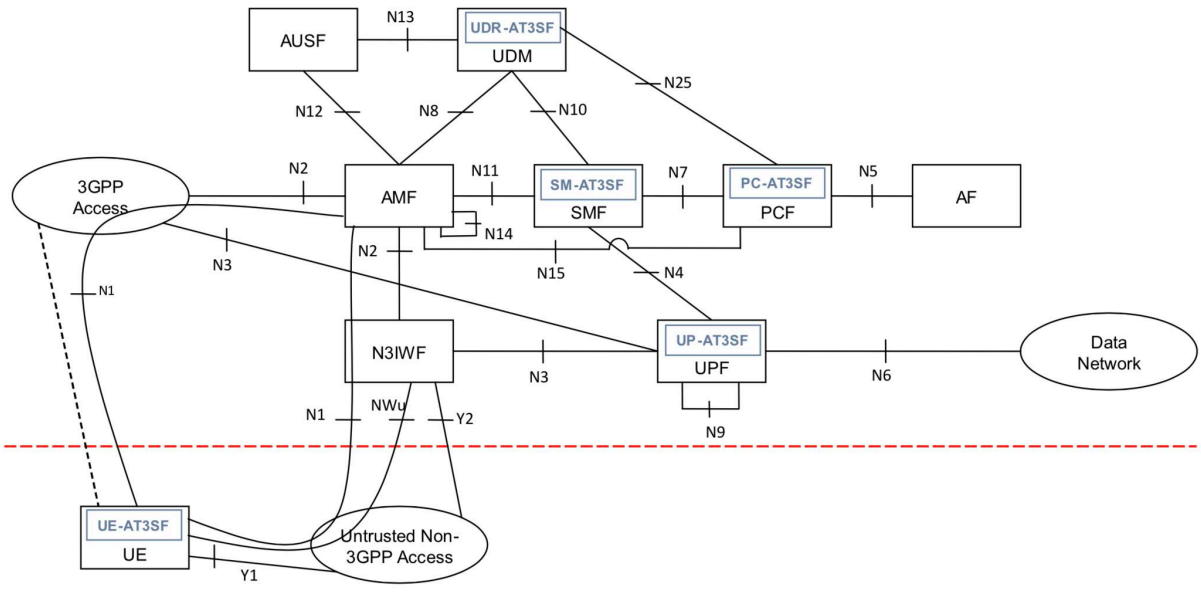


Figure 103: 3GPP's initial architecture for ATSSS

- User Data Repository for Access Traffic Steering Switching and Splitting Function (UDR-AT3SF): UDR-AT3SF holds UE ATSSS subscription data for operator service and user profiles. **This functional element is directly linked with the output of NWDAF-CEPE profiles, which receives and provides them –as will be described below- to other ATSSS functional elements.**

- Policy Control Access Traffic Steering Switching and Splitting Function (PC-AT3SF): **PC-AT3SF defines ATSSS policies** according to the application-specific information provided by the AF (via N5), access information/notification provided by the AMF (via N15), **UE ATSSS subscription and user profiles provided by the UDR-AT3SF** (via N25), network local policy or any combination of them. **The PC AT3SF may also take input from Network Data Analytics (NWDA) into consideration to generate or modify ATSSS policies.** The PC-AT3SF can provide ATSSS rules to UE-ATS3F.
- Control Plane Access Traffic Steering Switching and Splitting Function (CP-AT3SF): **CP-AT3SF is the main control plane of ATSSS architecture.** It is responsible for ATSSS policy enforcement and session management of all PDU sessions between 5GC and UE. CP-AT3SF can receive the ATSSS policies from PC-AT3SF via N7 and **generates ATSSS rules to control the behaviour of ATSSS** traffic by conveying ATSSS rules to UP-AT3SF over N4. CP-AT3SF may also receive access link information (e.g. access restriction, mobility status) from the AMF for all access legs as inputs to manage ATSSS behaviour. The CP- AT3SF can provide ATSSS PDU session related policies to UE-ATS3F during PDU session establishment and PDU session modification, as well as receive traffic usage reports from the UE and UP-AT3SF for dynamic ATSSS operations. **Based on the traffic usage reports, CP-AT3SF may send commands** (e.g., change access or access forbidden) to UE-AT3SF and UP-AT3SF via N1 and N4, respectively **to optimise ATSSS behaviour**.
- User Plane Access Traffic Steering Switching and Splitting Function (UP-AT3SF): UP AT3SF is the UP anchor point for all ATSSS traffic and presents a single IP address towards DN via N6. It is responsible for ATSSS policy rule enforcement in the UP of the core network and **relay traffic usage reports for from the UE** (if available) to CP-AT3SF via N4.
- UE Access Traffic Steering Switching and Splitting Function (UE-AT3SF): ATSSS policy rule enforcement at the UE for UE-initiated traffic (UL). It may also generate traffic reports to be sent to the CP-AT3SF. **The traffic usage reports from UE-AT3SF are directly linked to COmpAsS operation and fine tuning role, described in the previous section regarding the CEPE-COmpAsS complementarity.**

From the above analysis, it becomes clear that ATSSS' functional elements will be the responsible component/function, which will select the access technology per each active UE data flow. **3GPP is currently defining the context parameters, on which these decisions will be made, as well as the negotiation details between the UE and the network,** as far as the mapping between the data flows and the available access network resources are concerned. **COmpAsS, as already described in the respective sections, could operate as an ATSSS instance for this UE flow – RAT mapping optimisation, providing optimal selection with an energy- and signaling-efficient approach.** CEPE-NWDAF and COmpAsS are capable of providing this context information to the ATSSS elements. The next sub-section presents a holistic architecture, which integrates CEPE, COmpAsS (and CIP) and maps them with NWDA and ATSSS functionalities.

7.3.3 SDN-enabled, Cross Layer Control and E2E network slicing

In order to fully take advantage of the SDN-assisted approaches, in the forthcoming 5G architecture, such as network slicing, network data analytics methods –such as the knowledge extraction and profiling engine, CEPE, which was described earlier- will need to be exploited in order to optimize resource allocation and slice-flow mapping. NWDAF-CEPE's policies will, thus, need to be extended to the core network, applying policies to the SDN-enabled switches as well, formulating in an end to end manner network slices, adapted to the respective requirements.

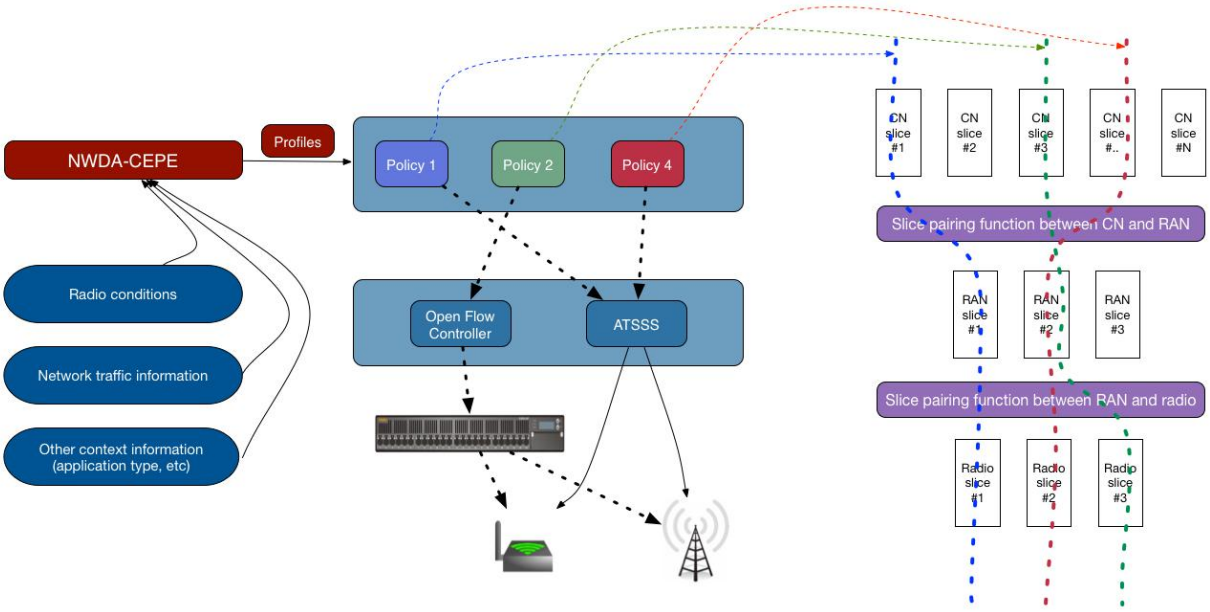


Figure 104: SDN-enabled Knowledge Extraction and Profiling

The figure above illustrates the afore-described concept: Besides NWDAF/CEPE's policies, which are generated for the optimal RAT and cell layer selection, the slice-related policies are being forwarded to NSSF, -via the Policy Control ATSSS Function- providing the overall resource allocation policy with the respective data path and available network slices optimal selection.

7.3.4 Integration with NWDAF, ATSSS in latest 5G architecture

According to the previous section, the Policy Control Access Traffic Steering Switching and Splitting Function **(**PC-AT3SF) defines ATSSS policies according to the application-specific information provided by the Application Function (AF), access information/notification provided by the Access and Mobility Function (AMF), and UE ATSSS subscription and user profiles provided by the User Data Repository ATSSS (UDR-AT3SF). **In our proposed architecture (Figure 105), this information –on which the ATSSS policies are generated- are federated by the NWDAF-CEPE function and are linked also to the NSSF, via the PC-AT3SF. COmpAsS module is considered part of the UE-AT3SF instance- and is responsible for sending traffic usage and access technologies Suitability reports to UP-AT3SF, which relays them "upwards" via CP-AT3SF ultimately to NWDAF-CEPE for feedback-based fine-tuning of policies. A COmpAsS instance is additionally deployed on the CP-AT3S function, in order to coordinate the scheme's reporting with the Control Plane.**

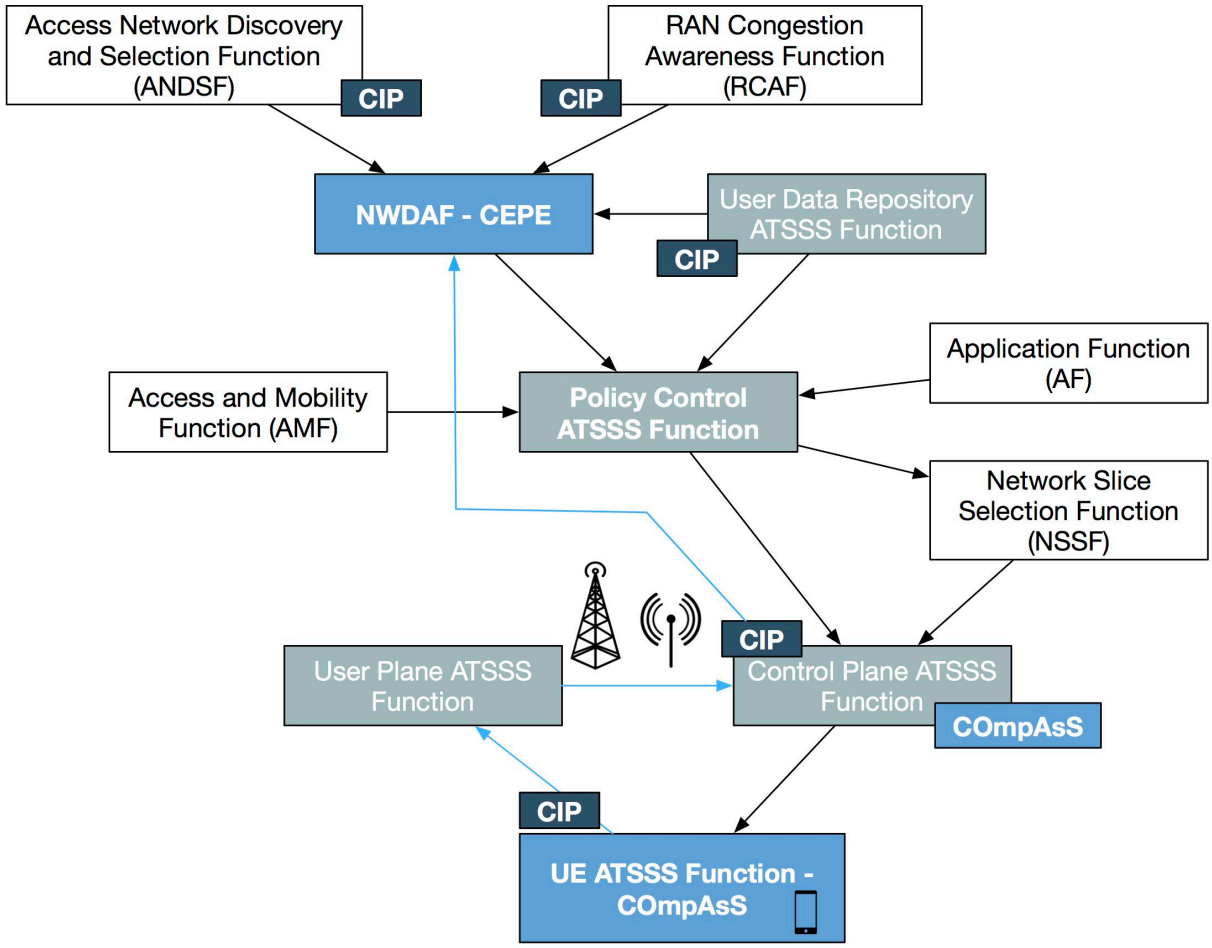


Figure 105: Logical interfaces of CEPE (as NWDAF instance) and COmpAsS (as UE-AT3SF instance) in 5G architecture

8. Supplementary Studies

In this section, some potential novel applications are introduced, which could take advantage of the afore-presented framework of context-based mechanisms, in order to address the forthcoming, challenging 5G network environments. Two studies are included in this section: a preliminary analysis on network traffic engineering approaches, which attempt to exploit CEPE and more specifically the user profiling methodology in order to apply holistic network resource planning policies, based on UE- and service-prediction mechanisms is the first; the second study focuses on a 5G use case application related to the Internet of Things -and more specifically, Precision Farming-, aiming to highlight explicit requirements related to industrial applications and 5G verticals overall, such as mission-critical machine type communication use cases, which are characterized by very high reliability and availability requirements, as well as very low end-to-end latency.

8.1 Traffic Engineering using CEPE

Traffic engineering is defined by the Internet Engineering Task Force (IETF) as the aspect of network engineering dealing with the issue of performance evaluation and performance optimization of IP networks [213]. Traffic Engineering must efficiently map traffic demands onto network resources and adaptively reconfigure the mapping to changing network conditions.

Traffic engineering methodologies are classified into two basic types: (i) time-dependent and (ii) state-dependent ([213]). In time-dependent traffic engineering, algorithms are used to optimize resource utilization in response to long time scale traffic variations (hours, days, weeks) using historical information on traffic patterns. State-dependent mechanisms deal with considerable variations in the actual/observed traffic load that could not be predicted using historical information and deal according to the current state of the network (traffic utilization, packet delay, packet loss, etc.).

In parallel, resource planning has a dual focus; resources and traffic. The resource-oriented aspect focuses on the optimization of network resources utilization in order to address future traffic demands. The traffic-oriented objectives aim at addressing the Quality of Service (QoS) requirements of the subscribers. Note that if these objectives are considered separately they lead to different solutions since the first focuses on the minimization of network resource usage while the second in providing the maximum possible QoS.

8.1.1 A quick insight on research efforts

The network dimensioning aspects of traffic engineering are discussed in details in ITU Teletraffic Engineering Handbook [217]. Specifically, in chapter 11, the notion of traffic matrix is introduced together with Kruithof's method for updates according to traffic forecasts. The latter is a typical approach used for choosing the optimal topology, (re)configuring a network and traffic routing.

In general, the key issue behind all traffic engineering efforts is the accurate estimation of a traffic matrix using load information. Towards this end, numerous approaches have been devised, broadly classified into two categories, namely gravity-based and tomography-based methods.

Gravity models apply Newton's law of gravitation on various scientific fields (i.e. social scientists model the movement of people between geographic areas, international trade of goods between countries is predicted using gravity models). In the case of traffic engineering, given two network nodes i and j the gravity model quantifies their traffic

matrix cell Xij as RiAj/fij9where Ri, is the incoming traffic to node i, Aj the outgoing traffic from node j (both equivalent to object masses) and fij their "friction" factor (equivalent to the masses' distance) derived using locality information. Typical examples appear in [223], [224], [225] and [226].

Tomography based methods operate based on the analogy of CAT scans. A CAT scan creates an holistic view of an organ or a body by concatenating consecutive scans; on the other hand network tomography considers traffic aggregation and attempts to find the quantities of interest, essentially solving the inverse problem. Simply stated, given the aggregated view Y of a network identified by routing matrix A we wish to find a matrix X (traffic matrix) for which XA=Y. Since Y is the aggregation of X under A it contains less known-quantities than the number of unknown-quantities in X. Methods that solve this problem are [227], [228], [230], [231].

A third approach is to consider the problem as a linear optimization one and employ polynomial time algorithms to solve it thus ensuring scalability (e.g. [232], [233]). The latter is important in case we apply the solution on a large scale environment. However this approach may exhibit two shortcomings; at first, we cannot always guarantee global optimization [234] while in parallel we cannot guarantee fairness in terms of load balancing between links (some links may be get overloaded).

8.1.2 Behavioral modeling and Forecasting with CEPE

The collection and subsequent processing of the information monitored through the lifetime of a network opens new horizons in the behavioral modeling of mobile subscribers. In the previous documents we showcased the exploitation of CEPE on an artificially generated dataset and quantified its added value through specific KPIs. Recall in parallel that CEPE comes with the sole prerequisite of mapping any incoming tuple to a vector format (i.e. transcending to a high dimensional space Rn). Therefore, it is not data model specific and may accommodate any kind of information.

Evidently we can exploit CEPE results in order to do more things than user behavioral profiling and customer segmentation. User profiling is a specific functionality and CEPE is a mechanism implementing it. Obtained results can then be exploited in numerous ways (e.g. in customer relation management, in operation support etc). In the context of the current research effort we proposed the derivation and dynamic update of a rule-set catering for the optimization of network operation.

However, the exploitation of the results can be manifold:

1. Employ the derived user and service profiles in order to predict service usage patterns per user class at a given location and time thus quantify the expected traffic per user class, service class, location, time etc.
2. Perform aggregations per service class (i.e. ignore user and device dimensions) and predict service usage patterns per time and location
3. Identify correlations between device classes and service consumption behaviors and in conjunction with open data related to the penetration of device types, forecast traffic evolution.
4. Exploit mobility information to perform profiling and prediction (mobility profiles per users as well as mobility patterns per type/class of user)

[9] The equation depicts a simple and general version of the gravity model.

5. Use derived information in conjunction with open datasets (e.g. related to demographics) in order to extract insights (for example, mobility patterns can be used for traffic management and urban planning).
6. Predict QoS/QoE per service, location and device type.
7. Predict the required resource reservation based on customer-flow profiling in order to address the capacity and QoS requirements
8. Predict the allocation of flows to nodes enabling traffic engineering

It is worth noting that most of the input information required in the aforementioned tasks is in the format of a timeseries, i.e. a sequence of data points, measured at successive points in time and spaced at uniform time intervals; it is therefore necessary to provide a small discussion regarding timeseries, their modeling and the way they are used for forecasting.

Formally, we denote as timeseries any set of values X={X1,X2,X3…} where Xi is a real number and i is a positive natural number denoting a specific instance of the time continuum. Typical tasks involve:

- Similarity Searching: Given a timeseries Q find its most similar timeseries P from a set of timeseries
- Dimensionality Reduction: Reduce the number of measurements while retaining information content
- Clustering: Identify groups of timeseries
- Classification: Assign a new timeseries to a group
- Prediction: Predict the behavior of a timeseries in the near future
- Modeling: How to properly model our data points so as to perform all aforementioned tasks in the easiest possible way.

An important factor that influences all issues discussed so far is the proper modeling of a timeseries. In other words, the way that we use to describe the procedure that generated our data. Having this description at hand we can easily predict future values, assess the similarity of two series, reduce their size by identifying periodicity and categorize them correctly. Due to the vast application field of timeseries (e.g. econometrics, weather, speech, etc.) there is a huge number of approaches and tools for modeling and prediction.

Usually the Autoregressive Model (AR), the Moving Average Model (MA), their combinations (ARMA and ARIMA - [214]) and the Box-Jenkins methodology for prediction [215] are used. In case of seasonality Seasonal ARIMA (SARIMA) and SARIMAX are employed -[214]. Moreover, Hidden Markov Models can also be applied for timeseries modeling and prediction. HMMs application is based on the assumption that the values of the series are generated by chain thus given a set of t values, we attempt to identify the states that generated it and forecast the t+1 value by considering which is the most probable transition state. Finally, Grey Box models ([216]), is another method which attempts to identify the governing laws of a timeseries and generate predictions.

8.1.3 CEPE-enabled Network Planning and Traffic Engineering

Building on the outcomes of a CEPE-enabled network, we can predict user behavior in terms of service consumption and therefore, derive the total traffic generated by specific service and user classes in a given geographic area or set of cells. Given the latter, we can then proceed and apply typical or custom traffic engineering solutions. In the context of this paragraph we provide some indications towards this direction.

Generally, as long as we can predict the expected traffic per user or user class we can apply typical approaches such as those discussed in ITU- E.760 [219]. ITU- E.490.1 [218], although dating back in 2003, accurately depicts the individual steps of traffic engineering –namely traffic demand (model, measure, forecast), service objectives (QoS requirements, objectives, network resources), traffic control and finally performance monitoring– thus highlighting the need for an extended dataset.

In parallel E.500 [222] contains hints and recommendations regarding grouping of measurements as well as classification of traffic (e.g. what should be considered normal/high/low, etc.). E.506 [220] discusses management of data in terms of missing values and prediction. It also provides an intuitive example of forecasting international traffic using 10 years of observations[10]. E.507 [221] explains the use of AR for forecasting international traffic and indicates how the Yule Walker equations can be employed for the calculation of AR coefficients.

Following the CEPE-enabled, rule-based, network optimization case presented in the context of the current contract we can design a similar framework for traffic engineering and resource planning. An adaptive, feedback-based control system that monitors the network, exploits historic information, builds profiles and forecasts KPI values can be combined with rules that focus on driving the network to a desired state by means of engineering traffic. It is evident however, that in order for any model to operate properly, an extended dataset should be considered incorporating for example geolocation information, CDRs/EDRs, etc.

The envisaged solution will comprise an integral part of an overall framework which will:

- Monitor (actively or passively) the network and quantify key variables that will be later exploited by CEPE (not a CEPE functionality)
- Extract profiles using the CEPE methodology (either supervised or unsupervised)
- Dynamically derive rules catering for short term network optimization. Update and enforce the set of rules using the methodology presented in SD5.
- Forecast future KPI values and traffic demands
- Map traffic demands onto network resources using any of the established approaches (e.g. linear programming, gravity models or tomography)

The novelty of the approach resides in the exploitation of the derived profiling scheme in order to perform forecasting and estimation of the overall traffic. This solution should be evaluated experimentally with SOTA approaches in order to assess to quality of the produced results. In parallel however, it should also be evaluated along the following axes:

- Scalability
    - An important property of the solution should be scalability so that it can scale to networks with hundreds to thousands of nodes. This property essentially implies that the solution should be fast and ideally exhibit a liner or polynomial time complexity.
- Avoid link overloading

[10] The specifications referenced in this paragraph are relatively old and focus on telephone network, yet they contain fundamental methods for data management and timeseries forecasting. In parallel, the methodology presented remains unaffected.

- Another aspect that should be taken into account is the capability of the algorithm to avoid link overloading (i.e., route traffic through specific links thus leading to their saturation). To cope with that issue the solution may restrict the maximum link load. Restricting link load has the advantage of enabling a linear programming formulation of the network dimensioning problem but may lead to sub-optimal resource allocation when the network is heavily loaded

- Ensure balancing of resources
  - Load balancing is guaranteed by means of distributing traffic by multiple paths based on the network state, responding to fast traffic variations and offering the potential of better network-wide load balancing.

In the previous section, we reported a series of experiments, which have been conducted using the NS-3 simulator and various network configurations. The aim of part of the past section's experiments was to investigate if and how the handover procedure is optimized when applying CEPE in the RAT selection related to the active UE flows; towards this end, in each experiment various measurements were extracted from the simulated network operating with and without CEPE and results are compared in order to establish the validity and viability of the concept. The evaluation was conducted via assessing pre-defined network KPIs, i.e., the throughput of the UEs, the experienced delay and the packet loss –both as average metrics for all users, as well as per service type-.

In this work, we performed a new series of experiments similar to the aforementioned simulations, but this time from the perspective of the overall system requirements. The goal of the profile-based traffic-engineering scheme is to identify context and profiling information that is valuable for resource planning and traffic flow characterization in 5G networks. To this end, key indicators are identified and used in order to capture the network resource planning and traffic flow requirements, e.g., capacity, throughput, application type, etc. In order to provide some initial insights, we use already-extracted profiles in the same way these were extracted for previous experiments. However, in next steps, a more specific profiling scheme must be developed that will take into account the overall system's KPIs and mainly focus on the profiles from the traffic engineering perspective.

The main target of this study is to identify the confidence in predicting the system's traffic flow requirements based on the application/service users' type extracted from CEPE's profiling scheme. This will ultimately lead in efficient resource allocation and spectrum management for the network operators towards the system's load balancing, as well as QoS optimization for the user.

The following table provides an overview of the different services that were configured during the simulations, along with the traffic parameters per service[11] in order to calculate the throughput requirements per type of service.

Table 24: Throughput requirements estimation per service

Service type	UL/DL	Parameters	Value	Throughput (Kbit/s)	Average service duration[3]
Voice	Uplink	Packet size (bytes)	18	14.4	1.8 minutes
		Interval (ms)	10		

[11] Values for the different service types are extracted based on references available in SD4, page 11

	Downlink	Packet size (bytes)	17	13.6	
		Interval (ms)	10		
VoIP	Uplink	Packet size (bytes)	40	32	20 minutes
		Interval (ms)	10		
	Downlink	Packet size (bytes)	60	48	
		Interval (ms)	10		
FTP	Uplink	Packet size (bytes)	12	0.192 (~0)	~10 seconds for 3 MBs file
		Interval (ms)	500		
	Downlink	Packet size (bytes)	600	2400 (2.4 Mbit/s)	
		Interval (ms)	2		
Web	Uplink	Packet size (bytes)	12	0.192 (~0)	4.2 seconds per page
		Interval (ms)	500		
	Downlink	Packet size (bytes)	800	3200 (3.2 Mbit/s)	
		Interval (ms)	2		
Video	Uplink	Packet size (bytes)	12	0.192 (~0)	Average YouTube video is 4.12 minutes
		Interval (ms)	500		
	Downlink	Packet size (bytes)	500	800	
		Interval (ms)	5		

According to [235], “network peak usage is as high as 45 service requests per UE per hour in peak busy hours”. In general, this statistic, -in combination with the different user types per scenario, as well as the average duration per service type-, should be used in order to calculate the expected (average) as well the maximum service requests for the particular simulation system, thus estimate the capacity requirements of the system. In such a case, we would assume that 45 service requests/UE/hour applies for all service types, i.e., Voice, VoIP, FTP, Web and Video.

However, in the particular simulations that we focus on, all users run the respective services throughout the whole simulation time (i.e., 600 seconds) according to the simulations’ configurations. In other words, the required throughput capacity is each moment calculated as the peak required system throughput, as all services are running simultaneously. Based on the type of the users, as a result, we predict the theoretical system throughput requirements. Then, the operator would “translate” the throughput

capacity requirements -taking into account the spectral efficiency- into bandwidth requirements.

In the simulation scenario we performed we evaluated 7 sub-scenarios, each time increasing the number of FTP users. In detail, the 7 sub-scenarios contain respectively [0, 8, 12, 16, 20, 24, 28, 40] FTP users (out of the 40 overall users). CEPE profiling enables us to create service profiles for each user and predict the types of applications that the user will consume when entering a specific network area. As a result, each moment, the network administrator is capable of having a holistic view of the theoretical network capacity in terms of throughput. Based on **Error! Reference source not found.** nd the respective values of uplink and downlink throughput, we calculate the estimated system capacity in terms of throughput in Table 25:

Table 25: User distribution per service and est. throughput capacity requirements

Sub-scenario	**User distribution per service type**	**Theoretical (peak) overall system throughput capacity requirement (Kbit/s)*** ***Calculated as (e.g. in sub-scenario 1 UL): 0 x video data rate + 10 x voice data rate +…+ 10 x VoIP data rate).**	
		Uplink	**Downlink**
1	0 FTP, 10 voice, 10 web, 10 video, 10 VoIP	467.84	40616
2	8 FTP, 8 voice, 8 web, 8 video, 8 VoIP	375.808	51692
3	12 FTP, 7 voice, 7 web, 7 video, 7 VoIP	329.792	57231.2
4	16 FTP, 6 voice, 6 web, 6 video, 6 VoIP	283.776	62769.6
5	20 FTP, 5 voice, 5 web, 5 video, 5 VoIP	237.76	68308
6	24 FTP, 4 voice, 4 web, 4 video, 4 VoIP	191.744	73846.4
7	28 FTP, 3 voice, 3 web, 3 video, 3 VoIP	145.728	79384.8

From the above table it becomes obvious that for the different users' distributions in each one of the sub-scenarios, the theoretical (peak) throughput capacity range for the uplink is between 145.728 and 467.84 Kbps, while for the downlink it is between 40.6 and 79.5 Mbps.

Apart from the theoretical (ideal) throughput requirement calculations, we predict the actual throughput based on historical measurements from the network. In the particular experiment, we perform different simulation runs in order to build our KPI database for our predictions. During these runs, on the one hand the mobility of the users varies, while on the other hand the service type of a single user may vary as well. The prediction, as a result, is performed on the basis of two variables: a) the varying mobility of the UEs, which is not defined in advance and b) the service type each UE will consume, based on the extracted CEPE profiles. In this way, we compare the predicted values (according to the CEPE-based profiling that links specific users to specific service types and –as a result- max. required throughput), with the actual result of one random experiment. The following figures (uplink and downlink) illustrate this comparison.

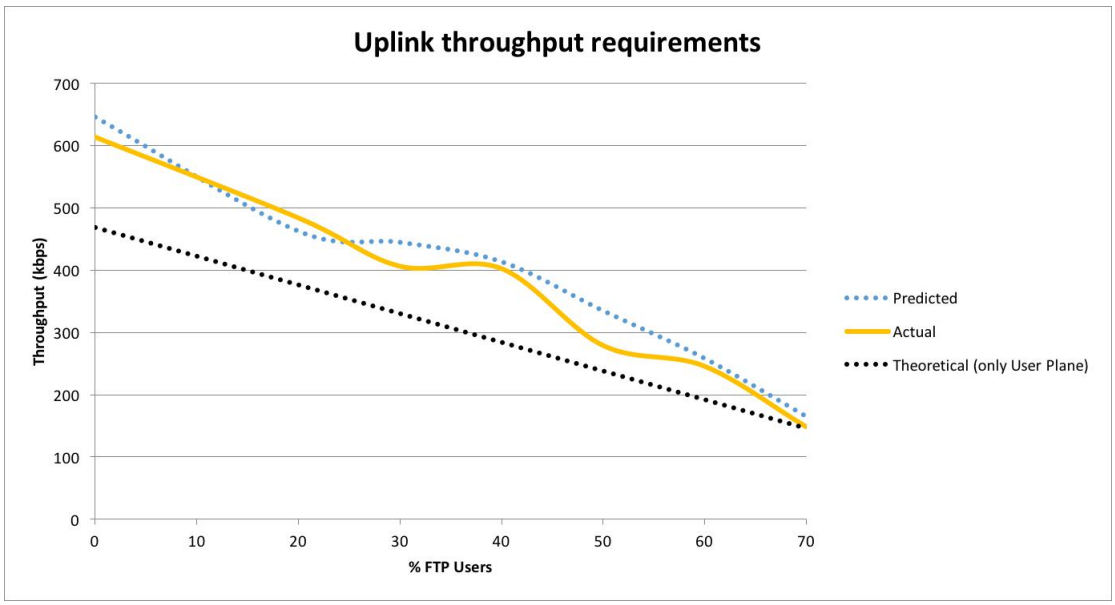


Figure 106: Theoretical, predicted and actual throughput requirements for uplink

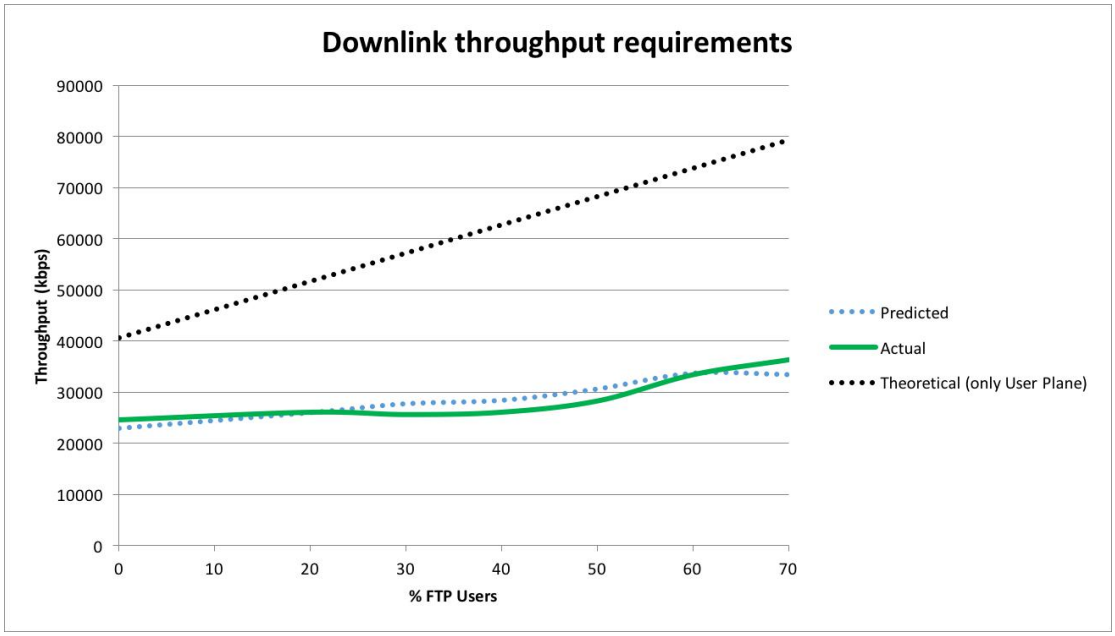
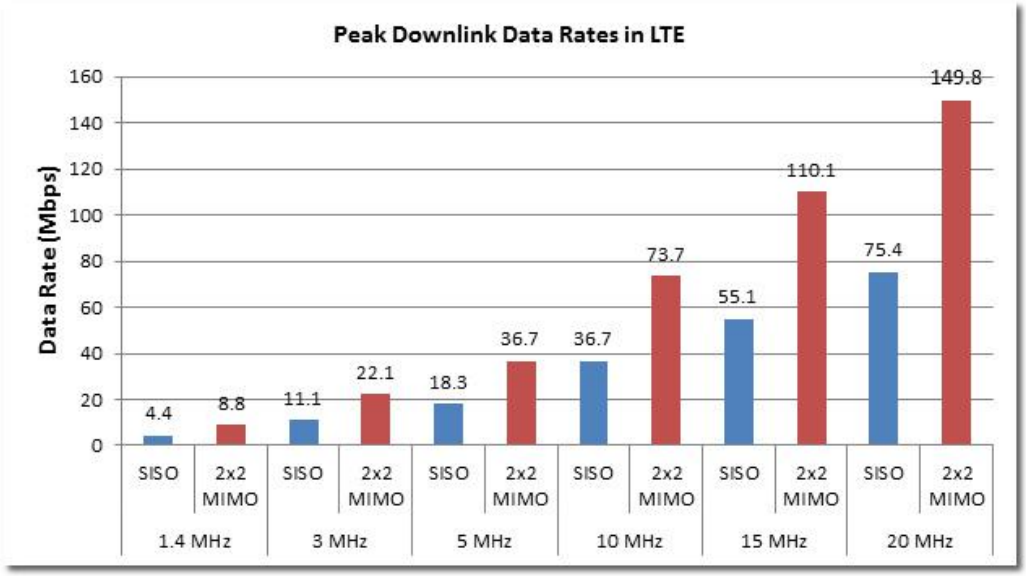


Figure 107: Theoretical, predicted and actual throughput requirements for downlink

The predicted values have been estimated based in historical information (multiple experiments on the same topology with random user mobility but same service consumption per user). In most cases, the predicted type of service was 100% successful. Nevertheless, minor divergences are observed in the sub-scenarios of 30, 40 and 50% of FTP users, with 5, 2.5 and 2.5% error in the service profile prediction accuracy. These minor errors contribute in the slight differences between the predicted and the actual results. All in all, the actual capacity requirements are almost identical to the predicted system's throughput variation, as the FTP users increase.

Throughput to bandwidth and vice-versa

The graphs in [236] illustrate how peak capacity varies according to the bandwidth and the modulation scheme. The generation of the graphs has been based on 3GPP document 36.213.



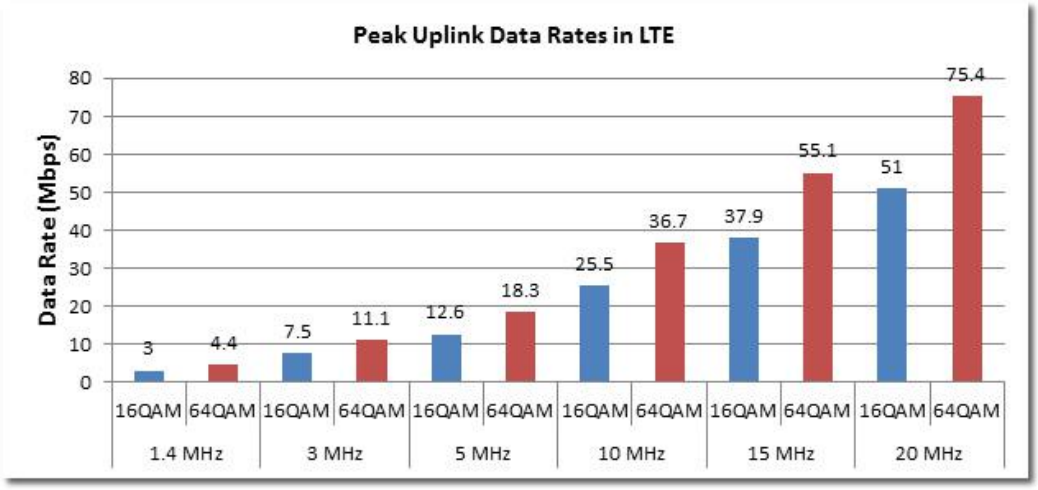


Figure 108a-b: Peak uplink and downlink capacity

Table 26: Resource Blocks – Bandwidth – Data rate

# of RBs	Bandwidth (MHz)	Downlink Data rate (Mbps)	Uplink Data rate (Mbps)
6	1.4	4.4 (SISO) / 8.8 (MIMO 2x2)	3/4.4
15	4	~11.1/22.1	7.5/11.1
25	5	~18.3/36.7	12.6/18.3
50	10	~36.7/73.7	25.5/36.7
75	15	~55.1/110.1	37.9/55.1
100	20	~75.4/149.8	51/75.4

According to the previous results, the operator –based on the predicted overall system's throughput requirements- makes the planning in terms of the spectrum to be utilized. Based on Figure 106 and Figure 107, as well as Table 26's mapping, the operator covers the traffic requirements of the particular scenario (topology, number of users, etc.) using

minimum spectrum for the Uplink, i.e., 1.4 MHz (or 6 RBs), and 5 MHz for the Downlink (MIMO) or 10MHz in case of SISO (25 or 50 resource blocks respectively).

8.2 5G Use Cases based on IoT and Context-based Network Slicing

An example related to a challenging Future Internet, Business-to-Business Collaboration Platform for Precision Farming

5G will support diverse vertical industries in a highly efficient manner, by the novel, advanced technologies it will comprise. These vertical industries introduce a variety of usage scenarios posing new challenging KPIs, some of the most common being very low latency, ultra high reliability, low energy consumption and support of massive connections.

In order to address those challenging KPIs, context-based radio resource management in 5G will highly rely on the Network Slicing concept [237], considered as one of the vital parts of the fifth generation networks' architecture. Network slicing will primarily address the deployment of multiple logical networks as independent business operations on a common physical infrastructure. The independent business operations will as a result utilize a portion of the network infrastructure and resources, which will be configured and optimized for the very specific requirements of the particular operation.

One of the relatively novel domains, which rely on a great extent upon the new technologies, which 5G will introduce, is the Internet of Things. IoT is mostly characterized by massive numbers of power-constrained devices, which infrequently transmit small amounts of delay-sensitive data. The type of communication, which characterizes such devices comprising the Internet of Things is Machine Type Communication (MTC); massive (mMTC) and ultra-reliable (uMTC) are two MTC sub-types, which are highly linked to the IoT requirements.

Smart Farming, is one of the IoT industries with high economic potential upon utilizing the 5G technologies. The first generation of Smart Farming applications concerned relatively simple applications with the involvement of a small number of monitoring sensors, which collect specific type of values; this information is aggregated in a database, residing close to the farm, and is being forwarded in a pre-defined time-interval (e.g., once a day) to a monitoring applications, providing an overview of the farm/greenhouse conditions to the app user/farmer. The fifth generation of wireless communications will offer a whole new variety of applications, with much more demanding, real-time tasks, which will often not involve human intervention. Massive MTC between sensors and actuators, computer-navigated machinery, etc., in massive deployments in farms, greenhouses, etc. will create a new landscape in the domain of Smart Farming industry. Towards, this direction, this section presents a brief overview of a Smart Farming system related to Smart Greenhouse Management and Control that was implemented in the context of the FIspace project ([238]) and utilizes the FIspace Business-to-Business platform. The figure that follows illustrates from a high level perspective the overall system (Figure 109).

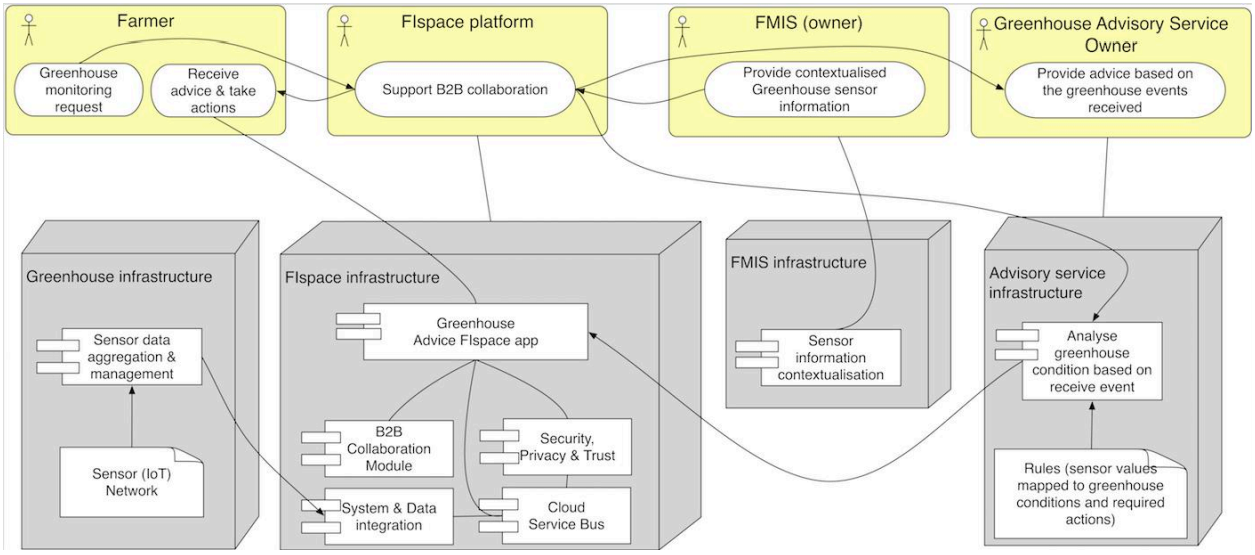


Figure 109: Smart Greenhouse Management and Control in 5G

The main use case is described by the following scenario: An IoT sensor network comprising several sensor types (temperature, humidity, luminosity, etc.) is installed inside a greenhouse, in order to monitor the greenhouse's conditions in real-time. Depending on the real-time status of the greenhouse conditions, actions can be taken by an automated system (actuators, e.g. watering machines, fire distinguishers, air condition systems, etc.) based on rules, which are applied via a Greenhouse Advisory Service. The final goal of the scenario is the optimization of the overall process and the maximization of the farm's/greenhouse's yield.

The different components, which comprise the overall system are typically located in different locations, as a result, -as also depicted in the above figure-, several connections and interfaces need to be supported, i.e., between the Greenhouse – where the IoT sensor network is deployed-, the FIspace infrastructure location, the Farm Management Information System (FMIS) location, as well as the Advisory Service infrastructure location. In cases of massive deployments, hundreds/thousands sensor devices from each farm are emitting small portions of critical data in a very frequent rate. Systems that are currently in place are failing because even in the areas that do have high speed connectivity they are failing due to high demand. According to the latest 5G use cases definition, massive IoT scenarios will be utilizing devoted network slices, which will be optimizing the specific traffic flow types addressing KPIs such as ultra high reliability and very low latency.

The network slicing could be realized either via a static approach, i.e. an operator provides a dedicated slice for the specific location and type of data. Another approach would be for the network to be capable of identifying in a more dynamic way such traffic type, and create the slice on the runtime. Such a mechanism would require a context-aware tool, able to identify the specific flows, or even predict their presence in advance, if based on analytics, such as CEPE. This leads to one of the future directions, we are also willing to take in relation to CEPE operation. Besides the traffic steering and RAT selection, slicing management could be directly linked with such a Context Extraction Engine and federate the overall slicing management architecture of the forthcoming 5G system, as already indicated in the previous section with 3GPP's latest Network Data Analytics module (NWDA) ([154]), which provides slice-specific network data analytics to the PCF.

9. CONCLUSIONS

5G network architecture and operation will require a holistic and comprehensive view of the network, in order to succeed in providing high quality services, making the most out of the scarce resource availability in the complex, heterogeneous, forthcoming, 5G radio environments. As it is thoroughly proved in the context of this thesis, context-awareness is of uttermost importance in order to acquire this holistic and comprehensive view of the network, -and by making intelligent decisions-, optimise the allocation of resources among the heterogeneous users, devices, things and access technologies. Towards maximizing the merits of context-aware systems' capabilities, novel technologies, which are introduced in the latest 5G-related standardization efforts should be also exploited, such as the Software Defined Networking concepts and approaches, Virtualized Network Functions, as well as Network Slicing.

The objective of this thesis is to analyze the concept of context-aware radio resource management (RRM) and traffic steering for 5G radio environments, and present solutions for these research areas, by also associating them with the latest 5G architecture insights. As already discussed, context awareness is the ability of the network elements –or computing systems- to acquire and reason about the context information and adapt accordingly the resource allocation policies according to the respective context. Context, is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves. RRM is the system level management of radio resources, co-channel interference, as well as other radio transmission characteristics in wireless communication systems, such as cellular networks, wireless local are networks (WLANs) and wireless sensor systems (WSNs).

Although numerous solutions have been proposed so far, which exploit diverse parts of the available context information items –related to the network, the user equipment, etc., this work has attempted to propose a novel method, which makes almost no assumptions; this work has attempted to evaluate from different aspects the proposed schemes in an environment, as realistic as possible. The innovation of this work is reinforced by the fact that the proposed schemes are based solely on assumptions in line with the latest 3GPP standardization efforts in terms of context information acquisition, attempting this way to highlight the realistic and viable aspect of the solutions for next generation wireless networks. To the best of the writer's knowledge, no research proposal has attempted to limit its assumptions totally in line with the standardization guidelines; on the contrary, the vast majority of solutions make numerous assumptions, which often lead non-realistic proposals.

Context awareness comes at a cost. The more information is acquired and processed, the higher the granularity of the context awareness, however the larger the burden, which is placed –both on the network, as well as the computing entities-; one of the crucial topics, which was analyzed in the context of this thesis was the signaling overhead evaluation for each one of the proposed mechanisms, as well as the type of the information acquired, which should be within the available information items and network entities, and in line with the latest standardization efforts towards 5G.

In the context of this thesis, the focus has been placed on three distinct context-aware mechanisms, which attempt to address the resource scarcity issues, which will be faced in the forthcoming ultra dense network deployments. All three mechanisms, focus on a different aspect of the context-aware approach: COmpAsS is a UE-based scheme, which attempts to select the most suitable radio access technology and point of access for the UE; CEPE, is an offline profiling engine, which is used to generate user and device profiles, and –based on these profiles- optimize the resource allocation- attempting to

map in an optimal manner the mapping between the available resources and user/device profiles; CIP –the 3rd core mechanism- is a context information pre-processing and filtering scheme, which acts in a complementary manner to the aforementioned schemes, targeting to minimize the signaling and processing burden, which is posed by the diverse and numerous context information items –especially in dense 5G deployments- with massive number of users, devices and coexisting access technologies.

This thesis provided a holistic study, which comprised comprehensive analyses from diverse aspects: architectural, algorithmic, experimental, analytical, etc. for all mechanisms. Besides the evaluation of the integrity and validity of each one of the three core schemes, a novel architecture is proposed, in line with the latest 3GPP standardization steps, comprising COmpAsS, CEPE and CIP as instances of the proposed novel network entities of the new 5G-EPC.

One of the crucial matters, on which focus was given, was the context acquisition process. In order to design COmpAsS's, as well as CEPE's system parameters, a comprehensive analysis on the network resources, respective interfaces and context information item types was made. In addition, an analytical approach was presented in the case of COmpAsS, which provided detailed insights on the information items, which are used, along with the signaling overhead required to aggregate them. To the best of our knowledge, there is no previous work, which attempts to quantify the signaling overhead of the proposed context-based mechanism, and juxtapose it with the gains measured in the network-related KPIs part.

The validity of each one of the mechanisms was showcased via an extensive set of experimental scenarios, carried out in line with the 5G Ultra Dense network scenarios and requirements, in realistic simulated topologies and with diverse access technologies and layers (macro cells, femto cells, Wi-Fi APs, etc.). The flexibility of the open source NS3 simulator, -which was used throughout all the experimentation-, enabled us to customize our environment according to very specific requirements, and –thus- achieve the realistic models we targeted.

This extensive demonstration proved a number of gains with regard to primary network KPIs, such as the maximization of the achieved throughput, the minimization of unnecessary handovers, as well the reduction of the latency measurements, particularly for delay-critical services, as described in the 5G verticals' requirements. The performance of the proposed schemes was juxtaposed to well established handover and RAT selection mechanisms, already deployed in 4G/LTE. This comparison highlighted numerous outcomes, both as far as the network and QoS KPIs are concerned (throughput, latency, packet loss, etc.), as well as the signaling overhead evaluation, since we compared with baseline –already deployed in the market- solutions, and not theoretical solutions found in the literature.

The innovation of the proposed framework is summed up by the following points:

- It comprises three distinct mechanisms, which are capable of operating both in a complementary, as well as a standalone manner, improving different aspects of the overall network operation. COmpAsS targets to improve the UE's QoS in terms of achieved throughout, latency and minimize number of unnecessary handovers, CEPE's primary target is to optimise the overall UE-flow mapping, with the available resources in an optimal manner from the overall network perspective, while CIP aims to optimise the context acquisition and transmission process for reducing the burden posed to the network.
- The evaluation of the proposed schemes was carried out in an as realistic as possible environment and with as few assumptions as possible with regard to the network environment, the system specifications, the UE traffic characteristics and

mobility, etc. In addition, the evaluation was made comparing the proposed schemes to well established, baseline solutions from the 4G/LTE systems, aiming to show the gains in a clear and easily reproducible way.

- The simulation was designed in line with the 5G verticals' specifications and use case requirements. Ultra dense deployments were used, with coexisting diverse radio access technologies, while high demanding applications were deployed in order to challenge the proposed schemes, in line with the very challenging 5G requirements.
- Thorough studies took place occasionally with regard to the latest 3GPP standardization steps in order to keep the proposed mechanisms as updated as possible, and close to the proposed architecture, interfaces and specifications of the 5G system use cases and requirements. Additionally, in the last part of this work, a direct mapping between the proposed schemes and the respective 5G network entities took place in the final section, as instances inside the 5G architecture.
- Unlike the vast majority of similar solutions, a great effort was given on the signaling cost analysis, which also juxtaposed its outcomes to the gains measured in the network-related KPIs part, for an overall assessment.

Overall, the research carried out in the context of this thesis proves that context aware systems are capable of improving remarkably the resource management in the challenging forthcoming 5G environments, especially when exploiting the ever increasing available context information from the various sources related to the user, the device, the traffic flow and the network. COmpAsS algorithm, along with the use of the light-weight approach of Fuzzy Logic on the UE side, results in great outcomes, as far as the achieved throughput, latency and packet loss are concerned, both in the downlink, as well as the uplink. UE profiling –as performed by CEPE- will be of uttermost significance in the complex, heterogeneous 5G environments, as it offers the capability to the network administrator to efficiently apply policies in order to optimise the mapping between UE profiles and network slices. Furthermore, the prediction capabilities of such a mechanism provide additional benefits in order to improve the resources distribution of the network. CEPE shows significant gains in terms of throughput and delay, being also capable –as shown earlier- of operating as a NWDAF instance in the context of the 5G system. As the context information, which is generated by the diverse network entities, rises dramatically, this thesis also concludes that a context information pre-processing mechanism is crucial in order to handle the great volumes of data, which are generated, in order to save both computing (e.g., related to analytics and profiling), as well as network/signaling resources.

Besides the innovations proposed in this thesis, there are of course still numerous aspects to be addressed and potential directions, towards which, an overall context aware system could be further optimised in order to fully support the 5th generation wireless and mobile systems. In the future we plan:

- To closely follow the next steps from 3GPP with regard to the NWDAF and ATSSS functionalities and perform the respective adaptations in relation to the interfaces, context information item types, etc. to CEPE and COmpAsS.
- To elaborate on the context abstraction layer, both from the operational as well as the architectural perspective, which will be acting on top of all 3 schemes. A novel algorithm will be implemented, which coordinates COmpAsS and CEPE operation in an optimal way, balancing the UE- and network-oriented decision making, depending on real-time, as well as historical/predicted context.
- New simulation scenarios to be designed and executed, which will integrate simultaneously all three mechanisms, while they will rely on a much more advanced implementation in terms of signaling.

- Further work on CIP distributed architecture and improve CIP's operation and communication between its numerous instances, in complex scenarios with high number of base stations, APs and UEs, where processing and network resources may be increased.

ACRONYMS

2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
ATSSS	Access Traffic Steering, Switching and Splitting
(C)AC	(Call) Admission Control
CPU	Central Processing Unit
CS	Cell (re-)Selection
D2D	Device to Device (Communication)
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
FIS	Fuzzy Inference System
FL	Fuzzy Logic
FLC	Fuzzy Logic Controller
GSM	Global System for Mobile Communications
HO	Handover
HSS	Home Subscriber Server
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IoT	Internet of Things
KDD	Knowledge Discovery in Databases
KPI	Key Performance Indicator
IMSI	International Mobile Subscriber Identity
ITU	International Telecommunications Union
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
LWA	LTE-WLAN Aggregation
M2M	Machine-to-Machine
MF	Membership Function
MME	Mobility Management Entity

mMTC	Massive Machine Type Communication
MNO	Mobile Network Operator
MTC	Machine Type Communication
NMS	Network Management System
NSSF	Network Slice Selection Function
NWDAF	Network Data Analytics Function
OFCS	Offline Charging System
PCRF	Policy and Charging Rules Function
PDCP	Packet Data Convergence Protocol
PER	Packet Error Rate
P-GW	Packet Gateway
PL	Packet Loss
QoE	Quality of Experience
QoS	Quality of Service
RAT	Radio Access Technology
RCAF	RAN Congestion Awareness Function
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
SC	Small Cell
SDN	Software Defined Networking
S-GW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
UE	User Equipment
UDN	Ultra Dense Network
UDP	User Datagram Protocol
uMTC	Ultra-reliable Machine-Type Communication
WLAN	Wireless Local Area Network

ANNEX

CEPE Unsupervised and Supervised implementation details

Table 27: Employed mathematical notation

Symbol	Explanation
$\boldsymbol{x}$	*Row vector x*
x_i	*The i-th coordinate of* $\boldsymbol{x}$
x	*Variable x*
x^t	*The value of variable x at time t*
$\boldsymbol{X}$	*Matrix X with observations as rows*
$\boldsymbol{x}^T$ *or* $\boldsymbol{X}^T$	*Transpose of* $\boldsymbol{x}$, $\boldsymbol{X}$
X_{ij}	*The value of cell i,j of* $\boldsymbol{X}$
$\boldsymbol{X}_{:,j}$ *or* $\boldsymbol{X}_{j,:}$	*The j-th column/row of matrix* $\boldsymbol{X}$
$\boldsymbol{X}_{1:n,:}$ *or* $\boldsymbol{X}_{:,1:n}$	*All rows/columns of* $\boldsymbol{X}$ *from index 1 to n*
$\lvert\boldsymbol{X}\rvert$	*The number of observations (rows) in* $\boldsymbol{X}$
$\boldsymbol{U, D, S, N}$	*The matrices of User, Device, Service and Network observations respectively*

We assume that all entities are represented as a high dimensional vector residing in R^n. Before delving into details, we provide a table summarizing the basic mathematical notation we employ. Note that our vectorization assumption renders the algorithm data agnostic and capable of accommodating changes in the underlying data model. Indeed, any change in the raw data necessitates an update in the mapping function (i.e. the function which will map an entity –for example a *User*– as a high dimensional point) leaving CEPE unchanged.

CEPE Unsupervised Version

The unsupervised version of CEPE operates under the salient assumption that there is some latent structure in the data collection (i.e. groups/clusters of similar observations) which attempts to unveil and formalize it through rules. The algorithm appears in Table 28.

We assume that the dataset is split into subsets in accordance with the variables in consideration. An intuitive selection could be the time and location axes (e.g. weekdays, 9:00-12:00 around coordinates {(x,y),(z,a)}) however this is indicative since the procedure can be applied along any data dimension. This step corresponds to descending from Level 0 to Level 1 as depicted in Fig 2.

Thereinafter, we break down every derived subset into entity specific data-chunks following a simple disaggregation approach; each observation is broken down into its constituting entities (e.g. User, Device, Network and Service chunks as depicted in Step 1 of Table 28).

Table 28: The unsupervised version of CEPE approach

Step	***function: unsupervised_CEPE***	
	Input Parameters	***X*****:** The dataset where rows are observations and columns are monitored variables
	u, d, s, n←number of user/ device/ service/ network-related variables in observations	
1.	$\boldsymbol{U\_X}_{:,1:u}$ $\boldsymbol{D\_X}_{:,u+1:u+d}$ $\boldsymbol{S\_X}_{:,u+d+1:u+d+s}$ $\boldsymbol{N\_X}_{:,u+d+s+1:\lvert X^T\rvert}$	Divide the dataset into entity specific chunks.
2.	***ul***,***Uc*** *_ spectral_clustering(**U**)* *ur _ decision_tree(**U**,**ul**)* ***dl***,***Dc*** *_ spectral_clustering(**D**)* *dr _ decision_tree(**D**,**dl**)* ***sl***,***Sc*** *_ spectral_clustering(**S**)* *sr _decision_tree(**S**,**sl**)* ***nl***,***Nc*** *_ spectral_clustering(**N**)* *nr _decision_tree(**N**,**nl**)*	Run spectral clustering and decision tree classification for all derived matrix chunks. Spectral clustering facilitates the identification of clusters (i.e. groups of similar observations) while the decision tree classifier extracts a rule-set which will be used for the classification/mapping of new observations to the identified clusters. ***ul*** denotes the label vector of user clustering (e.g. 'user_group_x', *ur* the classification rules (e.g. age>18 ^ income<50K → user_group_x) and ***Uc*** the corresponding centroids (e.g. average age 22.5, average income 35K).
3.	***v*** *_ distinct_classes(**ul** _ **dl** _ **sl** _ **nl**)*	Merge all label vector and identify the distinct elements (i.e. cluster labels) that will form the nodes of the graph. Store them into vector **v.**
4.	$E_{i,j}\_P(v_i\|v_j),\_ij$	Calculate the adjacency matrix
5.	***G*** *_ full_graph(**v**,**E**)*	Generate the graph
	Output Parameters	***G***: The graph *ur, dr, sr, nr*: The set of rules ***ul, dl, sl, nl***: The set of labels ***Uc, Dc, Sc, Nc***: The set of centroids

The next step entails the application of spectral clustering on the identified chunks and the derivation of entity-specific clusters (Step 2). We assume that the derived membership is correct and generate a knowledge model using a decision tree classifier. The tree model facilitates abstraction and generalization; the rules will help us categorize new instances faster without maintaining the whole dataset in memory or constantly updating the spectral decomposition.

Each distinct cluster label defines a node in a graph (Step 3), which is connected with another entity node via a weighted edge w_{ij}. The simplest way to weight edges is the use of conditional probabilities; for example, when connecting a user cluster with a service cluster, we can define a weight:

$$w_{i,j} = \frac{times\ that\ a\ user\ of\ this\ cluster\ consumed\ a\ service\ of\ that\ cluster}{number\ of\ services}.$$

The possible paths that traverse all entities (i.e. from a given *User* node to any *Network* node traversing all other entities) correspond to the different profiles (i.e. the different combinations of User – Device – Service – Network labels that CEPE identified in the data collection).

The spectral clustering algorithm appears in Table 29; at first we calculate the graph Laplacian L (Step 1) by taking into account the data pairwise similarities. Afterwards we derive its eigendecomposition (Step 2) and apply the eigengap heuristic on the matrix of eigenvalues **Σ** (**Σ** is a diagonal matrix with eigenvalues in descending order along its main diagonal – i.e. $\sigma_{ii}\neq 0$ and $\sigma_{ij}=0$ when $i\neq j$).

Spectral clustering is both a clustering and a dimensionality reduction algorithm. The new dataset is obtained by retaining the first *k* eigenvectors of **E** (Step 4) and embeds the original observations from R^n to R^k where n is the number of variables. We employ the k-means algorithm in order to discretize the result (i.e. derive the clusters from $\mathbf{X_{new}}$). Finally, taking into account the cluster membership information we calculate the centroids on the original matrix **X** (Step 6).

The computational complexity of the procedure is primarily dominated by the application of spectral clustering on matrixes **U**, **D**, **S** and **N** as well as the formation of graph **G** thus is upper bounded by $O(|\mathbf{v}|^2+|\mathbf{U}|^3+|S|^3+|\mathbf{D}|^3+|\mathbf{N}|^3)$. Memory requirements are upper bounded by $O(|\mathbf{v}|^2+|\mathbf{U}|^2+|S|^2+|\mathbf{D}|^2+|\mathbf{N}|^2)$ due to the storage of the graph **G** and the eigedecomposition of **U**, **D**, **S** and **N**.

Table 29: The spectral clustering algorithm

Step	***function: spectral_clustering***	
	Input Parameters	**X:** The dataset where rows are observations and columns are variables
1.	$\boldsymbol{L\_D - XX^T}$	Calculate the un-normalized graph Laplacian $\boldsymbol{L}$[12]. $\boldsymbol{D}$ is the degree matrix and $\boldsymbol{XX^T}$ the observations' pairwise similarities matrix according to their internal product (i.e. $\sum_{i=1}^{|x|} x_i y_i$ where **x**,**y** are observations –i.e. rows– of $\boldsymbol{X}$).
2.	$\boldsymbol{E, \_eig(L)}$	Derive the eigen-decomposition of $\boldsymbol{L}$
3.	*k _ eigengap* **(Σ)**	Use the eigengap heuristic and find the number of clusters *k.*
4.	$\boldsymbol{X_{new}\_E_{:,1:k}}$	Derive the new dataset.
5.	***labels*** *_ k-Means(*$\boldsymbol{X_{new}}$*,k)*	Run k-means clustering in order to discretize the result.
6.	***C*** *_ centroids(**X**,**labels**)*	Calculate the centroids on the original matrix $\boldsymbol{X}$.
	Output Parameters	***labels***: The labeling scheme as derived from k-means ***C***: The centroids calculated on the original matrix X.

CEPE Supervised Version

The key differentiating factor of the supervised case with respect to the unsupervised one is that we know in advance how we should break the data chunks into distinct clusters. The latter is due to the fact that we are aware of the underlying structure of our data collection.

[12] The normalized graph Laplacian can also be used as well as any scheme (e.g. kernel similarity, k-NN search etc) for the definition of the pairwise similarities matrix W. For simplicity reasons we present only the un-normalized graph Laplacian case together with the internal product similarity computation.

The formal description of the algorithm appears in Table 30. Again we assume that the dataset is split into subsets in accordance with the variables in consideration and we break down every derived subset into entity specific data-chunks as depicted in Step 1.

As soon as the pre-processing is performed, data is fed to a classifier (Step 2). Note that in general we can apply any type of classification algorithm. A Tree Classifier –e.g. the ID3 we used in the experiments of the unsupervised case– will produce a set of classification rules (i.e. *X & Y → Z*). A Lazy Learner –e.g. kNN, [196]– will construct a tree and then identify for each incoming observation its *k* closest instances. A multiclass SVM classifier –[196]– will produce, for each pair of classes, a set of support vectors which will define a linear equation that optimally separates the two classes. Evidently, any classifier is directly applicable without affecting the methodology.

The rest of the procedure is identical to the unsupervised CEPE. Conceptually, the approach is the same; the key differentiating factor, –which eventually provides better discrimination results–, is the existence of the correct labeling scheme, which significantly enhances and speeds up the procedure.

The computational complexity of the procedure is primarily dominated by the application of the decision tree classifier on matrixes **U**, **D**, **S** and **N** as well as the formation of graph **G** thus is upper bounded by O(|**v**|^2+u|**U**|log|**U**|+s|**S**|log|**S**| log|U|+d|**D**|log|**D**|+n|**N**|log|**N**|). Memory requirements are upper bounded by O (|**v**|^2 + |**X**|) due to the storage of the graph **G** and original data matrix.

Table 30: The supervised version of CEPE approach

Step	***function: supervised_CEPE***	
	Input Parameters	***X*****:** The dataset where rows are observations and columns are monitored variables ***ul, sl, dl, nl***: The labeling scheme
	u, d, s, n←number of user/ device/ service/ network-related variables in observations	
1.	$\boldsymbol{U\_X}_{:,1:u}$ $\boldsymbol{D\_X}_{:,u+1:u+d}$ $\boldsymbol{S\_X}_{:,u+d+1:u+d+s}$ $\boldsymbol{N\_X}_{:,u+d+s+1:\lvert X^T\rvert}$	Divide the dataset into entity specific chunks.
2.	*ur _ decision_tree(**U**,**ul**)* *dr _ decision_tree(**D**,**dl**)* *sr _decision_tree(**S**,**sl**)* *nr _decision_tree(**N**,**nl**)*	Run decision tree classification for all derived matrix chunks. *ur* denotes the classification rules (e.g. age>18 ^ income<50K → user_group_x).
3.	***v** _ distinct_classes(**ul** _ **dl** _ **sl** _ **nl**)*	Merge all label vector and identify the distinct elements (i.e. cluster labels) that will form the nodes of the graph. Store them into vector **v.**
4.	$E_{i,j}\_P(v_i \mid v_j), \_ ij$	Calculate the adjacency matrix
5.	***G** _ full_graph(**v,E**)*	Generate the graph
	Output Parameters	***G***: The graph *ur, dr, sr, nr*: The sets of classification rules

Rules Extraction and Feedback Loop

Until this point, CEPE description has focused solely on building and maintaining a knowledge base containing groups and associations between users, devices (e.g. type, capabilities, battery status, mobility, charging etc), RATs (e.g. type, cell-id, location) and Services. **We have stated however, that the goal is to map UEs to RATs and proposed to do so via a set of rules.** In the context of this paragraph, we will discuss the derivation of these rules from our knowledge model and their evaluation.

A first straightforward approach could be the manual extraction and evaluation of rules. A human expert (e.g. network administrator) identifies the various groups, studies their properties (e.g. group *A* is made up of prepaid subscribers that perform *x* top-ups a month and communicate primarily with SMS) and then derives the combinations that according to his expertise will optimize network operation. For example, consider an unsupervised case were we have identified the following profiles:

- User Group X: User Group 19-29, Prepaid subscription, 2GB available for data, when at 20% of credit, service consumption and calls drop sharply by 50%, moving at high velocity
- Service Group Y: Video Streaming (i.e. YouTube), VoIP Services (i.e. Skype) take up more than 80% of his time
- Device Group Z: Samsung Galaxy S4

A human administrator would probably come up with a rule that optimizes the QoE of the user class and avoids excessive network signaling, e.g., reduce the large number of HOs occurring due to high user velocity. A probable rule could be the following:

- User Group X ^ Service Group Y ^ Device Group Z → Macro Cell

This approach is plausible for an average number of groups/classes per entity (e.g. less than 10 per entity) considering that numerous cases can be grouped together. **But obviously, if we need finer granularity (i.e. higher level of detail per class) we need to come up with a semi-supervised or a totally unsupervised approach for rules derivation.**

Assuming that the network is configured to take optimal decisions most of the time, we can autonomously generate a set of rules, which upon application can ameliorate network conditions. The derivation is based on the graph constructed in the final step of CEPE. Recall that every node of the graph represents a distinct class of a particular and is connected with other nodes via weighted edges with edge weights depicting the probability of having instances of both classes on the same path.

We can employ two distinct strategies in order to derive the rule-set; directly apply Bayesian logic (i.e. Naïve Bayes classification), find all possible rules and rank them according to their score or alternatively identify the paths that traverse all class types (i.e. *User, Device, Service* and *Network*) and rank them according to the sum of weights.

This way, the case with the highest probability (i.e. appears most of time) is the one applied as a rule in similar situations. However, **such an approach is prone to mis-configurations; a correct but rare decision will be ignored and never be applied even when it should be. Evidently, in order to employ this procedure, we need a kind of feedback loop that will promote correct rules and degrade those invalid.**

Recall that in the beginning we proposed the extraction of a set of KPIs against which CEPE performance will be evaluated; depending on the induced amelioration or

deterioration on these KPIs we can create a score for each rule in the form $\frac{1}{n}\sum_{i=1}^{n} w_i p_i$ where w_i is an optional weight – importance – for the i-th KPI defined by an administrator, p_i the percentage change of the i-th KPI due to the application of the rule in question and *n* the number of evaluated KPIs.

The approach is depicted in Table 31. We assume that the network initiates with the rules derived according to the graph traversal plan and operates with these rules for a given time period *t*; afterwards, a set of predefined KPIs ***x*** is evaluated per user against their counterparts during operation period *t-1*.

For each set of *n* KPIs we calculate $\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i^{t-1}-x_i^t)}{x_i^{t-1}}$, where $p_i = \frac{(x_i^{t-1}-x_i^t)}{x_i^{t-1}}$, x_i^t the value of the i-th KPI obtained during trial period *t* and x_i^{t-1} the same value during the trial period *t-1* (or in the case that *t=1* when the system run without CEPE). **Depending on the importance of each KPI we can adjust the sum by including weights from a set *w* thus obtaining the final score for a given rule. Finally, we sum up the individual scores per user and obtain a holistic value for each rule or set of rules.**

Table 31: Evaluate a rule-set using feedback from all subscribers

Step	***function:* evaluate_rule_set**	
	Input Parameters	***r*^{t}:** A vector of matrixes containing as elements one matrix per KPI for evaluation time *t*. ***r*$^{t-1}$**: A vector of matrixes containing as elements one matrix per KPI for evaluation time *t-1*. ***w***: A vector containing the weights for each KPI.
1.	*r _ number of elements in* ***r***^{t}	Find the number of rules included for evaluation
2.	*score _ 0,* ***s*** *_ Ø*	
3.	*for i =1:1:r* $s_i = evaluate\_rule(r_i^t, r_i^{t-1}, w)$ *score +=* s_i	For all rules in this evaluation run, calculate the induced amelioration/ degradation percentage and add it to the overall score.
4.	$score = \frac{1}{r} score$	Normalize the score taking into account the number of rules
	Output Parameters	*score*: The score obtained for the particular rule set ***s***: A vector containing the scores for all rules. The i-th element of ***s*** contains the score of the i-th rule.

TABLE 32: Evaluating a single rule using feedback from all subscribers

Step	***function:* evaluate_rule**	
	Input Parameters	***X*^{t}:** A matrix containing as columns the values of the KPIs at time t for all users. ***X*$^{t-1}$**: A vector containing as columns the values of the KPIs at time t-1 for all users. ***w***: A vector containing the weights for each KPI.
1.	*u _ number of rows of* ***X***^{t} *n _ number of columns of* ***X***^{t}	Find the number of users and KPIs included in this evaluation
2.	*score _ 0*	

Step	*function:* **evaluate_rule**	
	Input Parameters	$\boldsymbol{X^t}$**:** A matrix containing as columns the values of the KPIs at time t for all users. $\boldsymbol{X^{t-1}}$: A vector containing as columns the values of the KPIs at time t-1 for all users. ***w***: A vector containing the weights for each KPI.
3.	*for i =1:1:u* $score += \frac{1}{n} \sum_{j=1}^{n} w_j \frac{(\boldsymbol{X}_{i,j}^{t-1} - \boldsymbol{X}_{i,j}^{t})}{\boldsymbol{X}_{i,j}^{t-1}}$	For all users in this evaluation run, calculate the induced amelioration/ degradation percentage and add it to the overall score.
4.	$score = \frac{1}{u} score$	Normalize the score taking into account the number of users
	Output Parameters	*score*: The score obtained for the particular rule

Querying a CEPE defined model

The algorithmic solutions presented in the previous sections create the knowledge base and associate rules set, upon which decisions will take place. In the context of this paragraph we will focus on the decision step; given a CEPE model and an observation, how to best assign it to a specific class and which rule should CEPE advice for invocation?

Recall that we have essentially structured a set of profiles and a set of rules so given an observation we want to identify the optimal set of classes (i.e. profile), on which it should be mapped and the proper rule to invoke.

Querying can take place either real-time or offline. Real-time search means that the required parameters will be periodically transmitted from the UE to the network, which in turn will feed them to the model and derive the classification of the observation.

Table 32: Querying a CEPE model

Step	*function: query_CEPE*	
	Input Parameters	***o*:** An observation ***ur, dr, sr*:** The sets of classification rules as derived from the decision tree classifiers of either supervised of unsupervised CEPE. *online/offline*: A flag signifying whether the procedure will run real-time of offline ***N***: The set of network mapping rules as derived from the graph traversal process
	u, d, s _ number of user/ device/ service variables in the given observation	
1.	$\boldsymbol{u\_o}_{1:u}$ $\boldsymbol{d\_o}_{u+1:u+d}$ $\boldsymbol{s\_o}_{u+d+1:\lvert o^T \rvert}$	Divide the observation into entities
	if online:	
2.	*ul _ classify(**u**,**ur**)* *dl _ classify(**d**,**dr**)* *sl _ classify (**s**,**sr**)*	Run the classifier for all derived observation chunks and attribute them the most fitting label (ul, dl, sl).
	else:	
3.	*ul, dl, sl _ get_historic(**u, d, s**)*	Search past observations

Step	***function: query_CEPE***	
	Input Parameters	***o*****:** An observation ***ur, dr, sr*****:** The sets of classification rules as derived from the decision tree classifiers of either supervised of unsupervised CEPE. *online/offline*: A flag signifying whether the procedure will run real-time of offline ***N***: The set of network mapping rules as derived from the graph traversal process
4.	***R*** *_ get_rules(ul, dl, sl,* ***N****)*	The set of rules accompanied by their score as derived by the graph traversal process.
	Output *Parameters*	R_1: The rule with the highest score

It now becomes apparent that the vertical division step is extremely helpful since static information (e.g. device capabilities, user preferences etc.) will not change thus classification will take place only once for the updated entities. As soon as the observation has been properly classified, we match it with the rules and apply the one with the highest rank. **In case of multiple matching rules we can randomly select and apply one since the subsequent evaluation step will assess its correctness.**

Offline querying on the other hand assumes that user behavior exhibits strong periodicity in terms of time and location. Therefore, the user profile and service profiles will not change over time enabling the exploitation of previous decisions.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021, url: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html [accessed July 2017]

[2] ITU report, ICT Facts and Figures 2016, url: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf [accessed July 2017]

[3] Henry Blodget, The Number Of Smartphones In Use Is About To Pass The Number Of PCs, https://www.businessinsider.com.au/number-of-smartphones-tablets-pcs-2013-12 [accessed July 2017]

[4] 5G Vision Brochure, 5G-PPP, The 5G Infrastructure Public Private Partnership: The next generation of communication networks and services

[5] S. F. Yunas, M. Valkama, and J. Niemelä, "Spectral and energy efficiency of ultra-dense networks under different deployment strategies," IEEE Communications Magazine, vol. 53, no. 1, pp. 90-100, 2015.

[6] B. Soret, K. I. Pedersen, N. T. K. Jørgensen, and V. Fernández-López, "Interference coordination for dense wireless networks," IEEE Communications Magazine, vol. 53, no. 1, pp. 102-109, 2015.

[7] 3GPP Specification Groups, http://www.3gpp.org/specifications-groups

[8] 3GPP TS 23.234, V13.1.0, 3 “GPP system to wireless local area network (WLAN) interworking; system description (release 13)”, March 2017

[9] 3GPP TS 23.401, V15.0.0, “GPRS enhancements for E-UTRAN access, (Release 12)”, June 2017

[10] 3GPP TS 23.402, V 14.4.0, “Architecture enhancements for non-3GPP accesses, (Release 12)”, June 2017

[11] 3GPP TS 24.312, V14.1.0, “Access Network Discovery and Selection Function (ANDSF) Management Objects (MO), (release 14)”, June 2017

[12] 3GPP TR 23.859, Local IP access (LIPA) mobility and Selected IP Traffic Offload (SIPTO) at the local network, Release 12, 04-2013

[13] Hotspot 2.0 (Release 2.0) Technical Specification, Wi-Fi Alliance, https://www.wi-fi.org/downloads-registered-guest/Hotspot_2-0_%2528R2%2529_Technical_Specification_Package_v1-2.zip/29728

[14] 3GPP, TR 23.865 “Study on Wireless Local Area Network (WLAN) network selection for 3GPP terminals; Stage 2”, September 2013

[15] BT & & Alcatel Lucent White paper, Wi-Fi Roaming building on ANDSF and HOTSPOT 2.0, October 2012

[16] Small Cell Forum, Integrated Femto-WiFi (IWF) Networks, March 2012, http://www.docstoc.com/docs/116290366/Integrated-Femto-WiFi- (IFW)-Networks

[17] 3GPP ETSI TS 122 129 V V7.0.0 (2006-03) - Technical Specification, Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Handover requirements between UTRAN and GERAN or other radio systems (3GPP TS 22.129 version 7.0.0 Release 7)

[18] 3GPP TS 36.304, “User Equipment (UE) procedures in idle mode (Release 12)”, v. 12.6.0, September 2015

[19] T. Qu, D. Xiao, D. Yang, W. Jin, Y. He., "Cell selection analysis in outdoor heterogeneous networks", Proc. of IEEE ICACTE, 2010

[20] Q.Tongwei, X.Dengkun, Y.Dongkai, "A novel cell selection method in heterogeneous LTE-advanced systems", ICBNMT 2010

[21] Z.Becvar, P. Mach, "Performance of fast cell selection in two-tier OFDMA networks with small cells", Wireless Days (WD), 2012 IFIP

[22] A.Daeinabi, K. Sandrasegaran, X.Zhu, "Performance evaluation of cell selection techniques for picocells in LTE-advanced networks", ECTI-CON 10th International Conference, 2013

[23] 3GPP “Heterogeneous Networks in LTE”, J. Wannstrom and Keith Mallinson, link: http://www.3gpp.org/technologies/keywords-acronyms/1576-hetnet (November 2015)

[24] R.Madan, J.Borran, Ashwin Sampath, N. Bhushan, A. Khandekar, Ji Tingfang, "Cell Association and Interference Coordination in Heterogeneous LTE-A Cellular Networks", Selected Areas in Communications, IEEE Journal, 2010

[25] J.Oh, Y.Han, "Cell Selection for Range Expansion with Almost Blank Subframe in Heterogeneous Networks," IEEE 23rd International Symposium on PIMRC, 2012

[26] H.S. Dhillon, J. G. Andrews, "Downlink rate distribution in hetnets cellular networks under generalized cell selection", IEEE Wireless Communication Letters, June 2013

[27] M.Bembe, Kim Jeongchan, T.Olwal, Han Youngnam, "Available bandwidth-aware cell selection for expanded regions of small cells adopting ABS", ICTC International Conference, 2013

[28] J. Sangiamwong , Y. Saito , N. Miki , T. Abe , S. Nagata, Y. Okumura, "Investigation on cell selection methods associated with inter-cell interference coordination in heterogeneous networks for LTE-advanced downlink", Proc. Eur. Wireless Conf., 2011
[29] S. Kim, Y. Lee, "Adaptive MIMO Mode and Fast Cell Selection with Interference Avoidance in Multi-cell Environments," 5th ICWMC, 2009
[30] Rahul Thakur, Sudeepta Mishra, C.Siva Ram Murthy, "A load-conscious cell selection scheme for femto-assisted cellular networks", PIMRC, IEEE 24th International Symposium, 2013
[31] M. Feng, X. She, L. Chen, and Y. Kishiyama, "Enhanced dynamic cell selection with muting scheme for DL CoMP in LTE-A", Proceedings of the IEEE Conference on Computer Communications (INFOCOM), 2010
[32] L. Gao, X. Wang, G. Sun and Y. Xu, "A Game Approach for cell Selection and Resource Allocation in Heterogeneous Wireless Networks", IEEE Conference SECON, 2011
[33] Han-Shin Jo, Young Jin Sang, Ping Xia, Jeffrey G. Andrews, "Heterogeneous Cellular Networks with Flexible Cell Association: A Comprehensive Downlink SINR Analysis", IEEE Transactions on Wireless Communications, Vol. 11, No. 10, October 2012
[34] European Patent Office, EP 2529581 A1, “Cell selection and reselection in a telecommunication network”
[35] European Patent Office, EP 2647246 A1, “Cell re-selection using a ranking algorithm”
[36] European Patent Office, EP 2292050 A2, “Cell selection and reselection in deployments with home NodeBs”
[37] European Patent Office, EP 2458919 B1, “Cell re-selection in a cellular telecommunications network”
[38] European Patent Office, EP 2387279 A1, “Cell (re)selection in a heterogeneous mobile communication system”
[39] European Patent Office, EP 2638734 A1, “Re-selection in heterogeneous wireless communication network”
[40] 3GPP TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Rel.14)”, V. 14.3.0, June 2017
[41] M. Ghaderi and R. Boutaba, “Call admission control in mobile cellular networks: A comprehensive survey," Wireless Communications and Mobile Computing, Vol. 6, No. 1, February 2006.
[42] M. Ahmed, “Call Admission control in wireless networks: A comprehensive survey”, Communications Surveys & Tutorials, IEEE (Volume 7, No1), 2005
[43] D. Niyato, E. Hossain, “Call admission control for QoS provisioning in 4G wireless networks: issues and approaches”, IEEE Network, Volume 19, Issue 5, 2005
[44] Hamid Beigy and M. R. Meybodi, “User Based Call Admission Control Policies for Cellular Mobile Systems: A Survey”, Elsevier Computer Communications, Vol.28, Issue 16, October 2005, pp. 1798 – 1813
[45] M. Jiang, M. Nikolic, S. Hardly and L. Trajkovic, “Impact of self-similarity on wireless data network performance”, in Proc. IEEE ICC’01, Helsinki, Finland, June 2001, pp. 477-481
[46] Ramesh Babu H.S., Gowrishankar, Satyanarayana P.S., “Call Admission Control Mechanism for optimal QoS in Next Generation Wireless Networks”, 2010 International Conference on Intelligent Systems, Modeling and Simulation
[47] Olabisi E. Falowo, H. Anthony Chan, “Joint Call Admission Control Algorithm for Fair Radio Resource Allocation in Heterogeneous Wireless Networks Supporting Heterogeneous Mobile Terminals”, IEEE CCNC 2010 proceedings
[48] Sabari Ganesh J., Bhuvaneswari P.T.V., “Enhanced Call Admission Control for WiMAX Networks”, IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011
[49] Seung-Que Lee, Ryu Byung Han, Nam-Hoon Park, “Call Admission Control for Hybrid Access Mode Femtocell System”, 4th IEEE International Workshop on Selected Topics in Mobile and Wireless Computing
[50] Sha Sha, Rosemary Halliwell, Performance Modeling and Analysis of Dynamic Class-Based Call Admission Control Algorithm using Fuzzy Logic for Heterogeneous Wireless Networks, 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11
[51] Cristian Olariu, John Fitzpatrick, Philip Perry and Liam Murphy, “A QoS based call admission control and resource allocation mechanism for LTE femtocell deployment”, The 9th Annual IEEE Consumer Communications and Networking Conference - Wireless Consumer Communication and Networking
[52] Shusmita A. Sharna and Manzur Murshed, “Impact on Vertical Handoff Decision Algorithm by the Network Call Admission Control Policy in Heterogeneous Wireless Networks”, 2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)]
[53] Younghyun Kim, Haneul Ko, Sangheon Pack, Wonjun Lee, and Xuemin (Sherman) Shen, “Mobility-Aware Call Admission Control Algorithm With Handoff Queue in Mobile Hotspots”, IEEE Transactions on vehicular technology, Vol. 62, No. 8, October 2013
[54] European Patent Office, EP 2497296 A1, “A method of call admission control for home femtocells”

[55] European Patent Office, EP 2055016 A2, “Backhaul-level call admission control for a wireless mesh network”
[56] European Patent Office, EP 2580929 A1, “Admission control for shared LTE network”
[57] European Patent Office, EP 1805938 A2, “Methods and systems for measurement-based call admission control in a media gateway”
[58] European Patent Office, EP 1661334 B1, “Call admission control system and method for interpreting signaling messages and controlling traffic load in internet protocol differentiated services networks”
[59] European Patent Office, EP 1641232 A1, “Call admission control in a VoIP network”
[60] European Patent Office, EP 2249544 A1, “Call admission control device and call admission control method”
[61] European Patent Office, EP 2317811 A1, “Call admission control device and call admission control method”
[62] European Patent Office, EP 2178326 B1, “Method for assigning bandwidth in the call admission control of lub interface”
[63] European Patent Office, EP 1796330 B1, “Call admission control device and call admission control method”
[64] European Patent Office, EP 1793639 B1, “Call admission control device and call admission control method”
[65] European Patent Office, EP 1510081 B1, “System and method for call admission control”
[66] European Patent Office, EP 1858286 A2, “System and method for call admission control”
[67] European Patent Office, EP 1231802 B1, “Apparatus for call admission control based on transmission power of base station”
[68] European Patent Office, EP 1715638 B1, “Method and apparatus for quality-of-service-based admission control”
[69] European Patent Office, EP 1796331 B1, “Apparatus and method for call admission control”
[70] European Patent Office, EP 2078350 B1, “Device and method for controlling overload”
[71] European Patent Office, EP 2076076 A1, “Method for performing service admission control”
[72] European Patent Office, EP 2076069 A1, “Method and system for performing service admission control”
[73] European Patent Office, EP 1079650 A2, “Call admission control method and apparatus”
[74] European Patent Office, EP 2208386 A1, “Method and apparatus for providing call admission control for voip over wireless local area networks using a transparent proxy agent”
[75] European Patent Office, EP 1486082 B1, “Method and system for performing call admission control in the uplink for third generation wireless communication systems”
[76] European Patent Office, EP 2106179 A1, “Call admission control method for preventing the congestion in the Abis interface in a mobile telecommunication system”
[77] European Patent Office, EP 1471764 B1, “Call-admission controller and method of call-admission control”
[78] European Patent Office, EP 1796413 B1, “Call admission control device and call admission control method in a wireless communication system”
[79] European Patent Office, EP 1671512 B1, “Adaptive call admission and policing control for a communication link with limited bandwidth”
[80] European Patent Office, EP 2053804 A3, “Radio controller, radio base station, radio communication system, call admission control method, program and recording medium”
[81] Magnus Olsson, Stefan Rommer, Catherine Mulligan, Shabnam Sultana, Lars Frid, “ SAE and the Evolved Packet Core: Driving The Mobile Broadband Revolution”, Sept. 2009
[82] 3GPP 23.829 – a01 V10.0.1, “Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)”, Release 10, October 2011
[83] 3GPP TS 23.234, V11.0.0, 3 “GPP system to wireless local area network (WLAN) interworking; system description (release 11)”, 09-2012
[84] 3GPP TS 24.312, “Access Network Discovery and Selection Function (ANDSF) Management Objects (MO), (release 12)”, June 2013
[85] Dionysis Xenakis, Nikos Passas, Lazaros Merakos, Christos Verikoukis, “Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms”, Communications Surveys & Tutorials, IEEE (Volume:16 , Issue: 1), 1st Quarter 2014
[86] G. Zhou, P. Legg and H. Gao, “A Network Controlled Handover Mechanism and its Optimization in LTE Heterogeneous Networks”, IEEE Wireless Communications and Networking Conference (WCNC) 2013
[87] Y.-H. Wang, G.-R. Huang, Y.C. Tung, “A handover prediction mechanism based on LTE-A UE history information”, 2014 International Conference on Computer, Information and Telecommunication Systems (CITS), 7-9 July 2014
[88] G. Yang, X. Wang, X. Chen, “Handover control for LTE femtocell networks,” 2011 IEEE Internat. Conf. on Electronics, Comm. and Control (ICECC), vol., no., pp.2670-2673, Sept. 2011.

[89] M. Z. Chowdhury, W. Ryu, E. Rhee, Y. M. Jang, “Handover between macrocell and femtocell for UMTS based networks,” IEEE 11th Internat. Conf. on Advanced Comm. Techn. (ICACT) 2009, vol.01, no., pp.237-241, Feb. 2009.
[90] J. Kim, T. Lee, “Handover in UMTS networks with hybrid access femtocells,” IEEE 12th Internat. Conf. on Advanced Comm. Techn. (ICACT) 2010, vol.1, no., pp.904-908, Feb. 2010.
[91] Z. Becvar, P.Mach, “Adaptive Hysteresis Margin for Handover in Femtocell Networks”, IEEE 6th Internat. Conf. on Wirel. and Mobile Comm., pp.256-261, Sept. 2010.
[92] H. Zhang, W. Ma, W. Li, W. Zheng, X. Wen, C. Jiang, "Signaling Cost Evaluation of Handover Management Schemes in LTE-Advanced Femtocell", 2011 IEEE 73rd Vehic. Techn. Conf. (VTC Spring), pp.1-5,May 2011.
[93] S. Wu, S. Lo, "Handover Scheme in LTE-based Networks with Hybrid Access Mode Femtocells", J. of Convergence Inform. Techn., vol. 6, no. 7, pp. 68-78, July 2011.
[94] B. Jeong, S. Shin, I. Jang, N. W. Sung, H. Yoon, "A Smart Handover Decision Algorithm Using Location Prediction for Hierarchical Macro/Femto-Cell Networks," 2011 IEEE Vehic. Techn. Conf. (VTC- Fall), pp.1-5, Sept. 2011.
[95] K. S. B. Reguiga, F. Mhiri, R. Bouallegue, “Handoff Management in Green Femtocell Network”, Internat. J. of Comp. Apps., vol. 27, no. 4, pp. 1-7, Aug. 2011.
[96] D. Xenakis, N. Passas, and C. Verikoukis, "A Novel Handover Decision Policy for Reducing Power Transmissions in the two-tier LTE network", 2012 IEEE Internat. Comm. Conf. (ICC), pp.1352-1356, June 2012.
[97] Dionysis Xenakis, Nikos Passas, Christos Verikoukis, “An energy-centric handover decision algorithm for the integrated LTE macrocell–femtocell network”, Computer Communications Elsevier Journal, Volume 35, Issue 14, 1 August 2012
[98] A. Ulvan, R. Bestak, M. Ulvan, "Handover Scenario and Procedure in LTE-based Femtocell Networks", The 4th Internat. Conf. on Mob. Ubiq. Comput., Syst., Serv. and Technolog., pp. 213-218, Oct. 2010.
[99] P. Xu, X. Fang, J. Yang, Y. Cui, "A User's State and SINR-Based Handoff Algorithm in Hierarchical Cell Networks", 2010 IEEE 6th Internat. Conf. on Wirel. Comm. Netw. and Mobile Comp. (WiCOM), pp.1-4, Sept. 2010.
[100] D. Lee, G. Gil, D. Kim, "A Cost-Based Adaptive Handover Hysteresis Scheme to Minimize the Handover Failure Rate in 3GPP LTE System", EURASIP J. on Wirel. Comm. and Netw., vol. 2010, no. 6, Feb. 2010.
[101] K. S. B. Reguiga, F. Mhiri, R. Bouallegue, "Handoff Management in Green Femtocell Network", Internat. J. of Comp. Apps., vol. 27, no. 4, pp. 1-7, Aug. 2011.
[102] Márquez-Barja, J. and Calafate, C.T. and Cano, J.-C. and Manzoni, P., “An overview of vertical handover techniques: Algorithms, protocols and tools”, Computer Communications, vol. 34, no. 8, pp. 985-997, ISSN 0140-3664, 2011.
[103] K. R. Rao, Zoran S. Bojkovic, Bojan M. Bakmaz, “Network Selection in Heterogeneous Environment: A Step toward Always Best Connected and Served”, Telsiks 2013, October 2013
[104] A. Ahmed, L. Merghem Boulahia, and D. Gaıti, “Enabling Vertical Handover Decisions in Heterogeneous Wireless NetworksQ A State-of-the-Art and a Classification” , IEEE Communications Surveys and Tutorials, accepted for publication. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6587998.
[105] P. Bellavista, A. Corradi, C. Gianneli, “A Unifying Perspective on Context-Aware Evaluation and Management of Heterogeneous Wireless Connectivity”, IEEE Communications Surveys and Tutorials, Vol. 13, No. 3, Third quarter 2011
[106] X. Yan, Y.A. Sekercioglu, and S. Narayanan, “A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks”, Computer Networks Vol. 54, pages 1848-1863, 2010
[107] Sunisa Kunarak, Raungrong Sulessathira, Eryk Dutkiewicz, “Vertical Handoff with Predictive RSS and Dwell Time”, TENCON 2013 - 2013 IEEE Region 10 Conference (31194)
[108] Nizar Zorba, Hossam Hassaneim, Christos Verikoukis, “Green Handover with a Hybrid Satisfaction Mechanism”, Proceedings of the Tenth International Symposium on Wireless Communication Systems - ISWCS 2013
[109] Fu-Min Chang, Hsiu-Lang Wang, Szu-Ying Hu, Shang-Juh Kao, “An Efficient Handover Mechanism by Adopting Direction Prediction and Adaptive Time-To-Trigger in LTE Networks”, Computational Science and Its Applications – ICCSA 2013, Lecture Notes in Computer Science Volume 7975, 2013, pp 270-280
[110] Panagis Magdalinos, Apostolos Kousaridas, Panagiotis Spapis, Giorgos Katsikas, Nancy Alonistioti, “Enhancing a Fuzzy Logic Inference Engine through Machine Learning for a Self- Managed Network”, ACM Springer Mobile Networks and Applications (MONET), MONAMI 2011
[111] L. Xia, L Jiang, and C. He, “A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method”, IEEE ICC 2007

[112] A. Kaloxylos, S. Barmpounakis, P. Spapis, N. Alonistioti, “An efficient RAT selection mechanism for 5G cellular networks”
[113] B. Ma, and X Liao, “Vertical Handoff Algorithm Based on Type-2 Fuzzy Logic in Heterogeneous Networks”, Journal of Software, Vol. 8, No 11, November 2013
[114] M. Zekri, B. Jouaber, and D. Zeghlache, “Context aware vertical handover decision making in heterogeneous wireless networks”, 6th IEEE International Workshop on Performance and Management of Wireless Mobile Networks
[115] World Intellectual Property Organization, WO2005025260 A1, “Mobility management in mobile networks based on context information”
[116] European Patent Office, EP2575392 A1, “Enhanced handover procedure in a E-UTRAN between HeNBs connected to a HeNB Gateway”
[117] US Patent Office, US20130273917 A1, “Enhanced handover procedure”
[118] World Intellectual Property Organization, WO2013150500 A3, “LTE to non-3GPP network traffic offload based on Quality of Service class indicator”
[119] US Patent Office, US8615241 B2, “Methods and apparatus for facilitating robust forward handover in long-term evolution (LTE) communication systems”
[120] Canada Patent Office, CA2680822 A1, “Mobility management (MM) and session management (SM) for SAE/LTE”
[121] US Patent Office, US20140003393 A1, “Method and apparatus for supporting handoff from GPRS/GERAN to LTE E-UTRAN”
[122] World Intellectual Property Organization, WO 2005071999 A1, “Call handover in a wireless local area network”
[123] World Intellectual Property Organization, WO 2003107704 A1, “Proactive deployment of decision mechanisms for optimal handover”
[124] US 20100150102 A1, “Adaptive handover mechanism for heterogeneous wireless network”
[125] European Patent Office, EP 2278840 B1, “Handover in a communication network comprising plural heterogeneous access networks”
[126] European Patent Office, EP 1670273 A1, “Handover of a mobile node between access networks of different technologies in a mobile IP telecommunications system”
[127] European Patent Office, EP 1915014 A2, “Method and apparatus for handover decision by using context information in a mobile communications network”
[128] World Intellectual Property Organization, WO 2006065019 A2, “Control method of handover traffic service in multi-cell environment and control apparatus thereof”
[129] 3GPP TS 29.060, V11.5.0, General Packet Radio Service (GPRS) Tunneling Protocol (GTP) across the Gn and Gp Interface, December 2012 Release 11.
[130] 4G Americas White Paper, Mobile Broadband Evolution Toward 5G: Rel. 12 & Rel.13 and Beyond, June 2015.
[131] Apple iOS Connection Manager, https://support.apple.com/en- us/HT20283 [accessed July 2017].
[132] IEEE 802.11u, IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan network- s-specific requirements, Amendment 9: Interworking with External Networks, 2011.
[133] 3GPP TS 23.852, Study on S2a Mobility Based on GPRS Tunneling Protocol (GTP) and Wireless Local Area Network (WLAN) Access to the Enhanced Packet Core (EPC) Network (SaMOG), September 2013.
[134] 3GPP TS 23.203 V13.7.0, Policy and Charging Control Architecture, March 2016 Release 13.
[135] 3GPP 24.237 V13.4.0, “Mobility between 3GPP Wireless Local Area Network (WLAN) interworking (I-WLAN) and 3GPP systems”, March 2016.
[136] 3GPP TS 24.302 V13.5.0, “Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks”, Release 13, March 2016.
[137] 3GPP TS 23.261 V13.0.0, “IP Flow Mobility and Seamless Wireless Local Area Network (WLAN) offload”, Release 13, March 2016.
[138] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler and Y. Koucheryavy, "5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends," in IEEE Journal on Selected Areas in Communications, vol. 33, no. 6, pp. 1224-1240, June 2015.
[139] M. Ayyash et al., "Coexistence of WiFi and LiFi toward 5G: concepts, opportunities, and challenges," in IEEE Communications Magazine, vol. 54, no. 2, pp. 64-71, February 2016.
[140] M. Ayyash et al., "Coexistence of WiFi and LiFi toward 5G: concepts, opportunities, and challenges," in IEEE Communications Magazine, vol. 54, no. 2, pp. 64-71, February 2016.
[141] Fraunfoher IPMS, Li-Fi, Optical Wireless Communication, http://www.ipms.fraunhofer.de/content/dam/ipms/common/products/WMS/WMS_OWC_2016_web.pdf [accessed 11/2016]

[142] X. Duan and X. Wang, "Authentication handover and privacy protection in 5G hetnets using software-defined networking," in IEEE Communications Magazine, vol. 53, no. 4, pp. 28-35, April 2015.
[143] H. Song, X. Fang and L. Yan, "Handover Scheme for 5G C/U Plane Split Heterogeneous Network in High-Speed Railway," in IEEE Transactions on Vehicular Technology, vol. 63, no. 9, pp. 4633-4646, Nov. 2014.
[144] J. S. Thainesh, N. Wang and R. Tafazolli, "A scalable architecture for handling control plane failures in heterogeneous networks," in IEEE Communications Magazine, vol. 54, no. 4, pp. 145-151, April 2016.
[145] Y. Kim et al., "Feasibility of Mobile Cellular Communications at Millimeter Wave Frequency," in IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 3, pp. 589-599, April 2016.
[146] H. Zhang, C. Jiang, J. Cheng and V. C. M. Leung, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," in IEEE Wireless Communications, vol. 22, no. 3, pp. 92-99, June 2015.
[147] H. Peng, Y. Xiao, Y. N. Ruyue and Y. Yifei, "Ultra dense network: Challenges, enabling technologies and new trends," in China Communications, vol. 13, no. 2, pp. 30-40, Feb. 2016.
[148] H. A. U. Mustafa, M. A. Imran, M. Z. Shakir, A. Imran and R. Tafazolli, "Separation Framework: An Enabler for Cooperative and D2D Communication for Future 5G Networks," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 419-445, Firstquarter 2016.
[149] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro and K. Zheng, "Toward 5G densenets: architectural advances for effective machine-type communications over femtocells," in IEEE Communications Magazine, vol. 53, no. 1, pp. 134-141, January 2015.
[150] A. Orsino, G. Araniti, A. Molinaro, A. Iera, “Effective RAT Selection Approach for 5G Dense Wireless Networks”, Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st.
[151] C.Y. Luiu, F.Y. Liu, A. Castiglionne, F. Palmieri, “Heterogeneous Network Handover Using 3GPP ANDSF”, 2015 IEEE 29th International Conference on Advanced Information Networking and Applications, March 2015, pp. 171-175
[152] D. Xenakis, N. Passas, L. Merakos, C. Verikoukis, “ANDSF-Assisted vertical handover decisions in the IEEE 802.11/LTE-Advanced network”, Elsevier Computer Networks Journal vol. 106, 2016, pp. 91-108.
[153] W. N. Schilit, "A system architecture for context-aware mobile computing", PhD Thesis, Columbia University, New York, 1995
[154] 3GPP TS 23.501, v 2.1.0, “System Architecture for the 5G System”, LTE Rel. 15, December 2017
[155] 3GPP TS 23.793, v 0.2.0, “Study on Access Traffic Steering, Switching and Splitting support in the 5G system architecture”, January 2018
[156] P. Makris, D.N. Skoutas, C. Skianis, "A Survey on Context-Aware Mobile and Wireless Networking: On Networking and Computing Environments’ Integration", IEEE Communications surveys and tutorials
[157] A. K. Dey, "Understanding and Using Context", Personal and Ubiquitous Computing, vol. 5, no. 1, pp. 4-7, 2001
[158] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen and M. Srivastava, "Using Mobile Phones to Determine Transportation Modes", ACM Trans. Sensor Networks, vol. 6, no. 2, pp. 1-27, 2010.
[159] Paolo Bellavista, Antonio Corradi, Carlo Giannelli, "A Unifying Perspective on Context-Aware Evaluation and Management of Heterogeneous Wireless Connectivity", IEEE Communications surveys and tutorials, Vol. 13, NO. 3, 2011
[160] Stenio Fernandes, Ahmed Karmouch, "Vertical Mobility Management Architectures in Wireless Networks: A Comprehensive Survey and Future Directions", IEEE Communications surveys and tutorials, Vol. 14, NO. 1, 2012
[161] D. Xenakis, N. Passas, L. D. Gregorio, and C. Verikoukis, "A context-aware vertical handover framework towards energy-efficiency", 73rd IEEE Vehicular Technology Conference (VTC), May 2011
[162] A. Klein, C. Mannweiler, J. Schneider, and H. D. Schotten, "A Concept for Context-Enhanced Heterogeneous Access Management", in Proceedings of the IEEE Global Communications Conference (GLOBECOM), 2010.
[163] C. Mannweiler, J. Schneider, A. Klein, H. D. Schotten, "From Context to Context-Awareness: Model-Based User Classification for Efficient Multicasting", Knowledge-Based and Intelligent Information and Engineering Systems, Lecture Notes in Computer Science Volume 6884, 2011, pp 146-154

[164] S. Carrella, C. Mannweiler, A. Klein, J. Schneider, H.D. Schotten, "A concept for context-aware multihoming with heterogeneous radio access technologies," Wireless Internet Conference (WICON), 2010 The 5th Annual ICST , vol., no., pp.1,7, 1-3 March 2010

[165] S. Mittal, A. Aggarwal, S.L. Maskara, "Application of Bayesian Belief Networks for context extraction from wireless sensors data," Advanced Communication Technology (ICACT), 2012 14th International Conference on , vol., no., pp.410,415, 19-22 Feb. 2012

[166] A. Abdallah, H. Frigui, P. Gader, "Context extraction for local fusion using fuzzy clustering and feature discrimination,", FUZZ-IEEE 2009. IEEE International Conference on Fuzzy Systems, vol., no., pp.490,495, 20-24 Aug. 2009

[167] R. Ocampo, L. Cheng, K. Jean, et al., "Towards a Context Monitoring System for Ambient Networks," Communications and Networking in China, 2006. ChinaCom '06. First International Conference on , vol., no., pp.1,3, 25-27 Oct. 2009

[168] S.S. Liao, J. W. He, T.H. Tang, “A framework for context information management”, JIS journal, vol.30, no.6, pp. 528-539, December 2004

[169] H. Peizhao, J. Indulska, R. Robinsion, “An Autonomic Context Management System for Pervasive Computing”, PerCom 2008, 6th Annual IEEE international Conference on Pervasive Computing and Communications, pp. 213-223, Hong Kong, March 2008

[170] K. Svanbro, J. Allden, “Communication System and Method for shared context compression”, Patent No US6950445 B2, Publication date: September 27th, 2005

[171] M. C. Chuah, “Header compression for general packet radio service tunneling protocol (GTP)-encapsulated packets”, Patent No US 6839339 B1, Publication date: January 4th, 2005

[172] M. Luna, M. Tervahauta, “Mobile network background traffic data management with optimized polling intervals”. Patent No US 8539040 B2, Publication date: September 17th, 2013

[173] A. Anand, J. Subtamanian, “Redundant traffic reduction in wireless networks”, Patent No WO 2014146755 A1, Publication date: September 25th, 2014

[174] G. S. Banavar, M. R. Ebling, G. D. H. Hunt, et al., “Method and Apparatus for Content Pre-Fetching and Preparation”, Patent No US 20090287750 A1, Publication date: November 19th, 2009

[175] 3GPP TS 23.861, V13.0.0, “Network-based IP Flow Mobility”, Release 13, June 2015.

[176] A. Kaloxylos, S. Barmpounakis, P. Spapis, N. Alonistioti, “An efficient RAT selection mechanism for 5G cellular networks”, International Wireless Communications and Mobile Computing Conference, 4-8 August 2014, Nicosia, Cyprus

[177] S. Barmpounakis, A. Kaloxylos, P. Spapis, N. Alonistioti, “COmpAsS: A Context-Aware, User-Oriented RAT Selection Mechanism in Heterogeneous Wireless Networks”, Mobility 2014, Fourth International Conference on Mobile Services, Resources, and Users, July 20-24 – 2014, Paris, France

[178] S. Barmpounakis, A. Kaloxylos, P. Spapis, N. Alonistioti, «Context-aware, user-driven, network-controlled RAT selection for 5G networks», The International Journal of Computer and Telecommunications Networking (COMNET Journal), Elsevier, Vol. 113, pp. 124-147, February 2017.

[179] L. Xia, L Jiang, and C. He, “A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method”, IEEE ICC, June 2007, pp. 5665 - 5670.

[180] B. Ma and X. Liao, “Vertical Handoff Algorithm Based on Type-2 Fuzzy Logic in Heterogeneous Networks”, Journal of Software, vol. 8, No 11, November 2013, pp. 2936-2942.

[181] 3GPP 36.304, V.13.1.0, “User Equipment (UE) procedures in idle mode”, Section 5.2.4.3 “Mobility states of a UE”, Release 13, March 2016.

[182] The METIS 2020 Project – Laying the foundation of 5G, www.metis2020.com, [accessed July 2017]

[183] Ns-3 simulator, http://www.nsnam.org/overview/what-is-ns-3/ [accessed July 2017].

[184] S. E. Elayoubi et al., “5G Service requirements and operational use cases: Analysis and METIS II Vision”, 2016 European Conference on Networks and Communications (EuCNC)

[185] 3GPP TR 36.931 v13.0.0, “Radio Frequency (RF) requirements for LTE Pico Node B”, Release 13, January 2016

[186] 3GPP TS 36.921 v13.0.0, “FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis”, Release 13, January 2016

[187] Traffic model related information, http://www.statista.com/statistics/185868/average-mobile-wireless-call-length-in-the-united-states-since-june-1993/

[188] Traffic model related information, http://en.wikipedia.org/wiki/Adaptive_Multi-Rate_Wideband

[189] Traffic model related information, http://networks.nokia.com/system/files/document/volte_white_paper_final.pdf
[190] Traffic model related information, http://www.swisscom.ch/dam/swisscom/en/res/mobile/mobile-network/netztest-connect-en-2014.pdfhttp://www.swisscom.ch/dam/swisscom/en/res/mobile/mobile-network/netztest-connect-en-2014.pdf
[191] Traffic model related information, http://www.theglobeandmail.com/technology/tech-news/how-much-bandwidth-does-streaming-use/article7365916
[192] Traffic model related information, http://www.sysomos.com/reports/youtube
[193] Traffic model related information, http://www.broadbandgenie.co.uk/mobilebroadband/help/mobile-broadband-usage-guide-what-can-you-get-for-your-gigabyte
[194] D. Drajic, S. Krco, I. Tomic et al., "Traffic generation application for simulating online games and M2M applications via wireless networks", 2012 9th Annual Conference on Wireless On-demand Network Systems and Services (WONS), IEEE, pp. 167-174
[195] P. Magdalinos, S. Barmpounakis, P. Spapis, A. Kaloxylos, et al., «A Context Extraction and Profiling Engine for 5G Network Resource Mapping», The International Journal for the Computer and Telecommunications Industry (COMCOM Journal), Elsevier Computer Communications, Volume 109, pp. 184-201, September 2017
[196] J. Han, M. Kamber, Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition. ISBN-9780123814791, 2011
[197] G. W. Stewart. "Matrix algorithms Vol I, II". Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.
[198] vonLuxburg, U. "A tutorial on spectral clustering". Statistics and Computing 17, 4, 395–416. 2007
[199] D. Mavroeidis, P.Magdalinos, "A Sequential Sampling Framework for Spectral k-Means based on Efficient Bootstrap Accuracy Estimations: Application to Distributed Clustering", ACM Transactions on Knowledge Discovery from Data, Volume 7, Issue 2, August 2012
[200] Bernd Theis, "The Great Mobile Network Test", Swisscom, available from http://www.swisscom.ch/dam/swisscom/en/res/mobile/mobile-network/netztest-connect-en-2014.pdf *[accessed April 2016]*
[201] WebSiteOptimization.com, http://www.websiteoptimization.com/speed/tweak/average-web-page/ [accessed April 2016]
[202] Tero Kuittinen, "Skype seems miles behind WhatsApp on daily engagement" available from http://bgr.com/2013/04/03/skype-2-billion-minutes-analysis-412280 *[accessed April 2016]*
[203] The NS3 network simulator*, https://www.nsnam.org/, [accessed August 2016]*
[204] Ericsson IP Network and Transport, Evolved Packet Core, SGSN-MME description, link: http://www.ericsson.com/ourportfolio/products/sgsn-mme?nav=productcategory004%7Cfgb_101_256, [accessed December 2015]
[205] 3GPP "Current Capacity of IMSI, MDISDN and IPv6 Identifiers", April 2010
[206] 3GPP TS 23.003, "Numbering, addressing and identification", v. 13.3.0, Rel.13, September 2013
[207] http://www.unixtimestamp.com/, [accessed December 2015]
[208] 3GPP TS 36.331, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification", v. 12.7.0, Rel.12, September 2015
[209] "eLTE2.2", Offline Charging Whitepaper, Huawei Technologies, February 2014
[210] 3GPP TR 23.803, "Technical Specification Group Services and System Aspects: Evolution of policy control and charging", v. 7.0.0, Rel. 7, October 2010
[211] 3GPP TS 29.212, "Policy and Charging Control (PCC)", v.11.6.0, Rel.11, October 2010
[212] 3GPP TS 25.331, "Radio Resource Control (RRC); Protocol Specification", v. 13.0.0, Rel.13, September 2015
[213] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. A framework for internet traffic engineering, November 2001
[214] Brockwell, P. J.; Davis, R. A. (2009). Time Series: Theory and Methods (2nd ed.). New York: Springer. p. 273. ISBN 978144190319
[215] Box, George; Jenkins, Gwilym M.; Reinsel, Gregory C. (1994). Time Series Analysis: Forecasting and Control (Third ed.). Prentice-Hall. ISBN 0130607746.

[216] Shimizu, N.; Ueno, O.; Komata, C., "Introduction of time series data analysis using grey system theory," Knowledge-Based Intelligent Electronic Systems, 1998. Proceedings KES '98. 1998 Second International Conference on , vol.2, no., pp.67,72 vol.2, 21-23 Apr 1998
[217] Teletraffic Engineering Handbook, ITU-D SG 2/16 & ITC, 2001, https://www.itu.int/ITU-D/study_groups/SGP_1998-2002/SG2/StudyQuestions/Question_16/RapporteursGroupDocs/teletraffic.pdf
[218] ITU E.490.1 – Overview of recommendations on traffic engineering
[219] ITU E.760 – Terminal mobility traffic modeling
[220] ITU E.506 (rev.1) – Forecasting International Traffic
[221] ITU E.507 – Forecasting International Traffic
[222] ITU E.500 – Traffic Intensity Measurement Principles
[223] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang, “Experience in measuring backbone traffic variability: Models, metrics, measurements and meaning”, in ACM SIGCOMM Internet Measurement Workshop, 2002.
[224] D. Lam, D. Cox, and J. Widom, “Teletraffic modeling for personal communications services,” IEEE Communications Magazine: Special Issues on Teletraffic Modeling Engineering and Management in Wireless and Broadband Networks, vo. 35, pp. 79–87
[225] J. Kowalski and B. Warfield, “Modeling traffic demand between nodes in a telecommunications network”, in ATNAC’95, 1995
[226] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, “Traffic matrix estimation: Existing techniques and new directions”, in ACM SIGCOMM, 2002
[227] Y. Vardi, “Network tomography: estimating source-destination traffic intensities from link data”, Journal of American Statistics Association, vol. 91, pp.365–377, 1996
[228] C. Tebaldi and M. West, “Bayesian inference on network traffic using link count data”, Journal of American Statistics Association, vol. 93, pp. 557–576, 1998
[229] J. Cao, D. Davis, S. V. Wiel, and B. Yu, “Time-varying network tomography”, Journal of American Statistics Association, vol. 95, pp. 1063–1075, 2000
[230] A. Adams, T. Bu, R. Caceres, N. Duffield, T. Friedman, J. Horowitz, F. L. Presti, S. Moon, V. Paxson, and D. Towsley, “The use of end-to-end multicast measurements for characterizing internal network behavior,” IEEE Communications Magazine, 2000
[231] M. Coates, et. al, “Internet tomography,” IEEE Signal Processing Magazine, 2002
[232] D. P. Bertsekas, “Network Optimization: Continuous and Discrete Models”, Athena Scientific, 1998
[233] T. Magnanti et. al, “Network Flows: Theory, Algorithms and Applications”, Prentice-Hall, 1993
[234] P. Aukia, M. Kodialam, P. Koppol, T. Lakshman, H. Sarin, and B. Suter, “Rates: A server for mpls traffic engineering” IEEE Network Magazine, vol 14, March 2000.
[235] “Managing LTE Core Network Signaling Traffic”, NOKIA, http://www2.alcatel-lucent.com/techzine/managing-lte-core-network-signaling-traffic/ [accessed September 2017]
[236] Frank Rayal Strategic insights & advisory in telecom and technology, http://frankrayal.com/2011/06/27/lte-peak-capacity/ [accessed September 2017]
[237] NGMN White Paper on 5G vision, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf [accessed September 2017]
[238] EU FIspace project, A Business-to-Business Collaboration Platform for Agri-food, Transport and Logistics, https://www.fispace.eu/