NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

DIVISION OF COMMUNICATION, ELECTRONIC AND INFORMATION ENGINEERING

Connectivity Management for HetNets based on the Principles of Autonomicity and Context-Awareness

DOCTORAL THESIS

of

Adamantia A. Stamou

Athens, October 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Διαχείριση της Συνδεσιμότητας για Ετερογενή Δίκτυα με βάση τα πρότυπα της Αυτονομικότητας και Επίγνωσης Περιβάλλοντος

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

της

Αδαμαντίας Α. Στάμου

Αθήνα, Οκτώβριος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Διαχείριση της Συνδεσιμότητας για Ετερογενή Δίκτυα με βάση τα πρότυπα της Αυτονομικότητας και Επίγνωσης Περιβάλλοντος

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

της

Αδαμαντίας Α. Στάμου

Συμβουλευτική Επιτροπή: Παπαβασιλείου Συμεών, Καθηγητής ΕΜΠ

Κοντοβασίλης Κίμων, Διευθυντής Ερευνών ΕΚΕΦΕ «Δημόκριτος»

Μάγκλαρης Βασίλειος, Καθηγητής ΕΜΠ

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 30η Οκτωβρίου 2019

……..	……..	……..
Παπαβασιλείου Συμεών Καθηγητής ΕΜΠ	Κοντοβασίλης Κίμων Δ/ντής Ερευνών ΕΚΕΦΕ «Δημόκριτος»	Ρουσσάκη Ιωάννα Επικ. Καθηγήτρια ΕΜΠ
……..	……..	……..
Κακλαμάνη Δήμητρα Καθηγήτρια ΕΜΠ	Μήτρου Νικόλαος Καθηγητής ΕΜΠ	Αλωνιστιώτη Αθανασία Επικ. Καθηγήτρια ΕΚΠΑ
	……..	
	Καρυώτης Βασίλειος Αναπλ. Καθηγητής Ιόνιο Πανεπιστήμιο	

Αθήνα, Οκτώβριος 2019

.....

Αδαμαντία Α. Στάμου

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Abstract

Within the Future Internet (FI) ecosystem, the Fifth Generation (5G) networks are already underway. These exploit higher frequency bands with wider available bandwidths and consider extreme base station and device densities, forming a Heterogeneous Network (HetNet) environment, aiming to meet the performance requirements of the lowest possible end-to-end latency and energy consumption. Efficient connectivity management in such a diverse networking environment is an open issue, towards attending user mobility between multiple Radio Access Technologies (RATs) and network tiers, confronting complexity and interoperability issues, accommodating application demands and user preferences and exploiting the capability of handling multiple active network interfaces concurrently. Collection, modeling, reasoning, and distribution of context in relation to sensor data would play a critical role in this challenge.

To this goal, the exploitation of the principles of context-awareness and autonomicity, should be exploited, as they enable the network entities to be aware of themselves and their environment, towards self-governing their behavior to achieve specific goals. Furthermore, proper assessment of the various VHO management approaches that present alternative context acquisition strategies, is needed, requiring a sufficiently comprehensive and generally applicable performance evaluation methodology, as the available methodologies for evaluating the performance of these proposals and for comparing alternatives are still limited.

Therefore, the contributions of this dissertation are twofold. The first part of the dissertation sheds new light to Vertical Handover (VHO) operations from an Autonomic Network Management (ANM) point of view, investigating the role of context-awareness

and self-x capabilities, by identifying the main concepts and providing a taxonomy of relevant architectural components and features, extending the current literature. Furthermore, representative state-of-the-art handover management solutions with context-aware and autonomic characteristics are presented, analyzed and correlated according to the proposed taxonomy and criteria, ultimately considering the overall enhancement of the VHO operations, culminating to conclusions that provide useful insights towards future, further enhanced solutions.

The second part of the dissertation provides a versatile modeling methodology, incorporating all significant effects that have an impact on performance, including signaling, processing and congestion (queuing theory). The resulting model is comprehensive, yet capable of admitting closed form solutions and can be flexibly tailored to different VHO architectures. To demonstrate the latter, we apply the modeling methodology in two context-aware VHO approaches that differ in the way of acquiring dynamically varying context (i.e. on-demand and proactively). For both approaches, the model-based results are validated against simulations, confirming the effectiveness and the accuracy of the modeling methodology, demonstrating that the proactive approach can provide significant delay and processing efficiency gains, leading in accordance, to potential energy consumption savings and lower OPEX and CAPEX costs.

Keywords: Autonomic Network Management, Context-Awareness, Cognition, Machine Learning, Proactive Computing, Connectivity (Mobility) Management, HetNets, 5G Networks, Future Internet, Modeling Methodology.

Περίληψη

Στο περιβάλλον του Διαδικτύου του Μέλλοντος, η Πέμπτη γενιά (5G) δικτύων έχει ήδη αρχίσει να καθιερώνεται. Τα δίκτυα 5G αξιοποιούν υψηλότερες συχνότητες παρέχοντας μεγαλύτερο εύρος ζώνης, ενώ υποστηρίζουν εξαιρετικά μεγάλη πυκνότητα σε σταθμούς βάσης και κινητές συσκευές, σχηματίζοντας ένα περιβάλλον ετερογενών δικτύων, το οποίο στοχεύει στο να καλυφθούν οι απαιτήσεις της απόδοσης ως προς την μικρότερη δυνατή συνολική χρονοκαθυστέρηση και κατανάλωση ενέργειας.

Η αποδοτική διαχείριση της συνδεσιμότητας σε ένα τόσο ετερογενές δικτυακό περιβάλλον αποτελεί ανοιχτό πρόβλημα, με σκοπό να υποστηρίζεται η κινητικότητα των χρηστών σε δίκτυα διαφορετικών τεχνολογιών και βαθμίδων, αντιμετωπίζοντας θέματα πολυπλοκότητας και διαλειτουργικότητας, υποστηρίζοντας τις απαιτήσεις των τρεχουσών εφαρμογών και των προτιμήσεων των χρηστών και διαχειρίζοντας ταυτόχρονα πολλαπλές δικτυακές διεπαφές. Η συλλογή, η μοντελοποίηση, η διεξαγωγή συμπερασμάτων και η κατανομή πληροφορίας περιεχομένου σε σχέση με δεδομένα αισθητήρων θα παίξουν κρίσιμο ρόλο σε αυτήν την πρόκληση.

Με βάση τα παραπάνω, κρίνεται σκόπιμη η αξιοποίηση των αρχών της επίγνωσης περιεχομένου και της αυτονομικότητας, καθώς επιτρέπουν στις δικτυακές οντότητες να είναι ενήμερες του εαυτού τους και του περιβάλλοντός τους, καθώς και να αυτοδιαχειρίζονται τις λειτουργίες τους ώστε να πετυχαίνουν συγκεκριμένους στόχους. Επιπλέον, χρειάζεται ακριβής ποσοτική αξιολόγηση της απόδοσης λύσεων διαχείρισης της συνδεσιμότητας για ετερογενή δίκτυα, οι οποίες παρουσιάζουν διαφορετικές στρατηγικές επίγνωσης περιβάλλοντος, απαιτώντας μια μεθοδολογία που να είναι περιεκτική και γενικά εφαρμόσιμη ώστε να καλύπτει διαφορετικές προσεγγίσεις, καθώς οι υπάρχουσες μεθοδολογίες στην βιβλιογραφία είναι σχετικά περιορισμένες.

Το σύνολο της μελέτης επικεντρώνεται σε δύο θεματικούς άξονες. Στο πρώτο θεματικό μέρος της διατριβής, αναλύεται ο ρόλος της επίγνωσης περιβάλλοντος και της αυτονομικότητας, σε σχέση με την διαχείριση της συνδεσιμότητας, αναπτύσσοντας ένα πλαίσιο ταξινόμησης και κατηγοριοποίησης, επεκτείνοντας την τρέχουσα βιβλιογραφία. Με βάση το προαναφερθέν πλαίσιο, ταξινομήθηκαν και αξιολογήθηκαν λύσεις για την υποστήριξη της κινητικότητας σε ετερογενή δίκτυα, οι οποίες δύνανται να θεωρηθούν ότι παρουσιάζουν επίγνωση περιβάλλοντος και αυτο-διαχειριστικά χαρακτηριστικά. Επιπλέον, μελετήθηκε κατά πόσον οι αποφάσεις που λαμβάνονται ως προς την επιλογή του κατάλληλου δικτύου, σύμφωνα με την κάθε λύση, είναι αποτελεσματικές και προτάθηκαν τρόποι βελτιστοποίησης των υπαρχουσών αρχιτεκτονικών, καθώς και προτάσεων προς περαιτέρω ανάπτυξη σχετικών μελλοντικών λύσεων.

Στο δεύτερο θεματικό μέρος της διατριβής, αναπτύχθηκε μια ευέλικτη αναλυτική μεθοδολογία, περιλαμβάνοντας όλους τους παράγοντες που μπορούν να συνεισφέρουν στην συνολική χρονοκαθυστέρηση, λαμβάνοντας υπόψιν την σηματοδοσία, την επεξεργαστική επιβάρυνση και την συμφόρηση (μελέτη ουράς), επεκτείνοντας την τρέχουσα βιβλιογραφία. Η μεθοδολογία είναι περιεκτική, ενώ ταυτόχρονα προσφέρει κλειστού τύπου λύσεις και έχει την δυνατότητα να προσαρμόζεται σε διαφορετικές προσεγγίσεις. Προς απόδειξη αυτού, εφαρμόσαμε την μεθοδολογία σε δύο λύσεις με διαφορετική στρατηγική επίγνωσης περιβάλλοντος (μια μεταδραστική και μια προδραστική). Και για τις δύο προσεγγίσεις, τα αναλυτικά αποτελέσματα επιβεβαιώθηκαν από προσομοιώσεις, επιβεβαιώνοντας την αποτελεσματικότητα και την ακρίβεια της αναλυτικής μεθοδολογίας. Επιπλέον, αποδείχθηκε ότι η προδραστική προσέγγιση εμφανίζει καλύτερη απόδοση ως προς την συνολική χρονοκαθυστέρηση, ενώ χρειάζεται σημαντικά λιγότερους επεξεργαστικούς πόρους, παρουσιάζοντας πιθανά

οφέλη και στην συνολική ενεργειακή κατανάλωση και στα λειτουργικά και κεφαλαιουχικά κόστη (OPEX και CAPEX).

Λέξεις Κλειδιά: Αυτονομικότητα, Επίγνωση Περιβάλλοντος, Μηχανική Μάθηση, Προδραστική Προσέγγιση, Διαχείριση της Συνδεσιμότητας, Διαχείριση της Κινητικότητας, Ετερογενή Δίκτυα, Δίκτυα 5^{ης} Γενιάς, Διαδίκτυο του Μέλλοντος, Γενική Αναλυτική Μεθοδολογία.

Acknowledgements

My pursue towards the PhD made me think about the role of research in the global knowledge economy context and the role of PhD students as future leaders and innovators, and also as mentors of younger generations. In this context, a European-wide framework for doctoral education, I think is essential to be established. I continue to reflect on the research processes optimization, towards continuous self- and team improvement.

I would like to thank my supervisor Symeon Papavassiliou for his guidance and support and especially for his understanding in the difficulties that a PhD student faces, both emotional and financial, and always managing to do the best for all his PhD students. I am also grateful for the trust to me and all his research team that enables us to flourish and lead our own way of doing things, cultivating us as future leaders.

I would like to thank my supervisor Kimon Kontovasilis for his guidance and support and the fact that took me -as he does with any of his PhD students- under his guidance, personally, insisting in a way of academic excellence. I am also grateful to him for believing in me since even before my graduation of my first degree.

I would also like to thank Nikos Dimitriou, for his guidance and support and his attitude of treating me as his colleague, since the first time we started cooperating, being an undergraduate student to today. I also thank Vasileios Karyotis for his excellent cooperation and his paradigm on team leadership.

In order to reach the point of starting a PhD, several people have to inspire you and show you the way. Among those people, there were the supervisors of my diploma thesis, Christos Verikoukis, Jesus Alonso Zarate and Luis Alonso, who made me realize that I want to pursue a research career. I am also grateful to Sotiria Psoma, Angela Moneda and Malamati Louta for their educational paradigm during my undergraduate studies.

Lastly and foremost, I thank my parents for being the first and greatest teachers for me, during all my life.

To all colleagues that pursue their PhD in Greece.

Contents

List of Figures

List of Tables

List of Abbreviations

Symbol	Description
ANDSF	Access Network Discovery and Selection Function
APAV	Access Point Acceptance Value
APSV	Access Point Satisfaction Value
ABC	Always Best Connected
ABS	Always-Best-Satisfying
AHP	Analytic Hierarchy Process
AHM	Autonomic Handover Manager
ANM	Autonomic Network Management
BER	Bit Error Rate
CNAPT	Client Network Address and Port Translator
CPN	Conventional Parameter Normalization
DF	Decision Function
DM	Deterministic Markovian
EPC	Evolved Packet Core
5G	Fifth Generation
FSA	Finite State Automata
TFFST	Finite State Transducer with Tautness Functions and Identities
FI	Future Internet
FL	Fuzzy Logic
GRA	Grey Relational Analysis
HO	Handover
IS	Information Server
MDP	Markov Decision Process
MIIS	Media Independent Information Service

MAC	Medium Access Control
MN	Mobile Node
MAD	Multiple Attribute Decision
MEC	Mobile Edge Cloud
MIH	Media Independent Handover
NFV	Network Function Virtualization
NN	Neural Networks
OSS	Operations and Support Systems
PB	Policy-Based
QoS	Quality of Service
RSS	Received Signal Strength
RTT	Round-Trip-Time
SNAPT	Server Network Address and Port Translator
SINR	Signal to Noise plus Interference Ratio
SAW	Simple Additive Weighting
SDN	Software Defined Networks
TOPSIS	Techniques for Order Preferences by Similarity to Ideal Solution
VHO	Vertical Handover

1. Introduction

The provision of ubiquitous broadband network access for mobile users has been a key research issue for a number of years, promoting the notion of a FI environment, consisting of open, intelligent and collaborative wireless and wire-line access networks [**1**]. Within the FI ecosystem, the *Fifth Generation (5G)* networks are already underway. These exploit higher frequency bands with wider available bandwidths and consider extreme base station and device densities, forming a *Heterogeneous Network (HetNet)* ecosystem, targeting at the lowest possible *energy consumption* and *end-to-end latency* [**2**], while catering to different service requirements. In this complex network ecosystem, macro cells will coexist with small cells (such as fempto and pico cells) utilizing multiple radio access technologies (RATs) [**3**].

In parallel, the FI vision includes the Internet of Things, regarding the management of information about real world objects and their surroundings, provided by an enormous number of sensors, wireless communications devices and embedded systems operating in different environments and providing a number of different services [**4**], [**5**]. In such a heterogeneous and complex networking ecosystem, users should be able to have contextualized, proactive and personalized access to services everywhere, under a seamless experience [**6**], extending the 'always best connected' (ABC) notion.

Therefore, it becomes essential to take a unified approach that integrates all diverse networking technologies available [**6**], towards enabling seamless roaming between networks, while accessing applications with different service requirements, and towards providing enhanced Quality of Service (QoS) and user satisfaction. Collection, modeling, reasoning, and distribution of context in relation to sensor data would play a critical role in this challenge [**7**]. Hence, a *context-aware* Vertical Handover (VHO) management framework is needed, in order to choose optimally the appropriate time to initiate the handover and the

most suitable access network for each specific service, to ensure service continuity and robustness against link and network impairments.

In this direction, the ideal answer and at the same time, the key, to ensure seamless connectivity in a complex heterogeneous FI environment could be provided by the vision of ANM, encompassing *context-awareness*, *self-management* and *cognitive* functionalities. ANM addresses the ability of networks to be aware of themselves and their environment and self-govern their behavior to achieve specific goals [**8**], without compromising the performance of the other coexisting networks or the global network performance metrics. ANM shares motivation and has confluent goals with other emerging technologies, such as Software Defined Networks (SDN) and Network Function Virtualization (NFV), as all three concepts seek to increase the flexibility, reliability and efficiency of operations and optimize network management and control. As it has been recognized, the notions of ANM, SDN and NFV can coexist [**9**], [**10**], [**11**]. In particular, ANM could be used to promote the local optimum in balance with the global optimum and the self-awareness of each distributed entity could be used to build the global awareness, enabling the development of appropriate global policies used to optimize the operation of the whole network.

Furthermore, various architectures and frameworks have been proposed, considering the management of VHOs in a heterogeneous set of Radio Access Networks (RANs), with the aim of enabling effective, context-aware network selections, but the available methodologies for evaluating the performance of these proposals and for comparing alternatives are still limited. Major standardization activities have provided specifications, notably the IEEE 802.21 [**12**] (and its evolutions IEEE 802.21 2017 [**13**] and 802.21.1 [**14**]) and the 3GPP ANDSF [**15**], for a unified VHO management framework. Part of these specifications addresses the collection and exchange of context related to the Candidate access Networks (CN) or the User Equipment (UE, also called Mobile Node (MN) in the following). The

aforementioned standards-based frameworks provide support primarily for static, time-invariant, context (e.g., a list of RANs serving a given area) and the architecture of these frameworks includes a Context Server (CS), serving as a repository of the relevant information.

In addition to these facilities, however, dynamically varying context is also necessary. In particular, resources availability context (depending on the current network loading conditions) is crucial for an optimal network selection, towards avoiding issues such as the occurrence of a series of unnecessary handovers (ping-pong effect) and bringing to the user the desired quality QoS. Due to the static nature of the information stored in the CS of the standards-based VHO frameworks, the acquisition of dynamic resource availability context occurs reactively, on an on-demand basis: Each time a handover is triggered, the MN subject to handover (or its Serving Network (SN)) exchanges signaling messages with the CNs, to determine the resources availability status therein. Several proposals have been made, as amendments or extensions of the standards-based VHO frameworks, including [**16**], [**17**], [**18**], [**19**], either equipping the original CS with the capability of receiving and storing dynamic context, or including one or more additional CS for the dynamic context.

Proper assessment of the relative merits of alternatives such as those just mentioned requires a sufficiently comprehensive and generally applicable performance evaluation methodology. However, most quantitative assessments that exist in literature rarely illustrate the exact process and the sources of the context-related information dissemination, and they result to a simple methodology demonstrating the steps of the VHO process and a basic signaling message exchange, adding up constant times. Such methodologies overlook various important complexities such as queueing phenomena. In addition, there are several brute-force simulation studies specific to a proposed framework, where it is hard to extrapolate the analytical methodology behind them and use the results for other approaches in order to

compare and contrast them, in order to prove the feasibility of each approach for efficient target network selection in next generation HetNets.

1.1 Contribution

On the first part of the dissertation, the field of autonomic VHO management is surveyed, by employing concepts of ANM to VHO management for the first time, in order to shed new light to VHO operations from an ANM point of view, investigating the role of context-awareness and self-x capabilities, towards encompassing FI environments and the emerging 5G networks [**20**].

A number of earlier studies (including [**21**], [**22**], [**23**], [**8**], [**24**], [**25**] and [**26**], among others) have surveyed issues related to ANM in general, but without specializing on VHO management, while other studies (e.g., [**27**], [**28**], [**29**], [**30**], [**6**], [**31**]) have focused only on general aspects of VHO management. Additionally, publications [**32**] and [**33**] have surveyed several purposed VHO management solutions featuring some degree of intelligence or a cognition potential. Despite such prior works, however, to the best of our knowledge there is a lack of a study that focuses particularly on *autonomic VHO management* in the FI era, defining the subject and providing a comprehensive analysis.

We start by reviewing basic concepts regarding cognition and autonomicity. Subsequently, we employ these concepts in an analysis of the autonomic handover management procedures under the light of the autonomic functions *monitor, analyze, plan,* and *execute*, providing an overview of the involved sub-processes and corresponding algorithms. We introduce a new taxonomy of the relevant architectural components, considering the scenario of context-aware MNs that operate within a complex FI environment and self-manage their mobility behavior.

Building upon the aforementioned taxonomy, we proceed towards addressing another issue of paramount importance for autonomic systems, namely *self-optimization*. A number of

important *autonomic features* related to autonomic handover management are discussed, each one promoting the system's self-optimization along a certain direction, towards the overall enhancement of the VHO operations. In connection with the autonomic features mentioned, we also investigate *robustness* issues associated with the VHO parameters and metrics of the network selection decision function. Such considerations relate to the ability of a system to achieve stable decisions under conditions of partial and possibly imprecise knowledge of contextual information, still a largely open issue in the present state of the art on network selection frameworks [**33**], [**6**].

Furthermore, on the second part of the dissertation, we provide a modeling analysis methodology, focusing on signaling in the VHO preparation phase, which incorporates all significant aspects associated with the exchange and processing of the signaling messages among the relevant architectural components, including the exact process of how context (including dynamic resource information) is made available [**34**]. The aim is to investigate the impact in delay-related performance, including all the involved transmission, processing and waiting delays. The system model is comprehensive, yet able to produce closed form results. The generic modeling analysis methodology is flexible, especially designed to adapt to different architectures that present different strategies of checking the resource-related information.

More specifically, we present a standard-based reactive approach, which can be defined as an on-demand resource information gathering approach. Through the proposed methodology, it can be illustrated how architectural choices affect the congestion in terms of their major architectural components. It can also be demonstrated the impact of computational resources scaling, in view of the overall end-to-end delay, in order to prove the feasibility of the approach for efficient RAN selection in next generation HetNets. Analytical results are verified by extensive simulation results, under an appropriately rich set of relevant

parameters. The proposed generic analytical methodology is properly designed to be applicable in relative heterogeneous network scenarios that are going to be presented at the next step of the doctoral thesis, combining the principles of context-awareness and autonomicity.

1.2 Structure

The general concepts of context-awareness, cognition and autonomicity are introduced in Chapter 2, followed by the proposed taxonomy and classification framework for context-aware VHO management in view of ANM, filling the existing literature gap. Proposed related autonomic features and robustness considerations for context-aware connectivity management are presented in Chapter 3. Accordingly, the developed concepts considering the proposed taxonomy and classification are applied to representative state-of-the-art context-aware handover management solutions with autonomic characteristics. Specifically, Chapter 4 reviews key characteristics of selected autonomic connectivity management solutions, providing a comparison of these solutions according to the framework, and presenting useful insights towards future, further enhanced solutions.

Furthermore, Chapter 5 presents the main elements of the proposed quantitative modeling methodology for context-aware connectivity management in HetNets, after presenting the necessary background of the major VHO management frameworks and related work. With the appropriate application of these generic methodology steps, the model can be adapted for the evaluation of different architectural approaches, as it is illustrated in Chapter 6, presenting an on-demand context-aware approach, and in Chapter 7, presenting a proactive context-aware approach. Chapter 8 provides numerical and also simulation results, showing the validation of the proposed analytical modeling methodology and illustrating the performance of the presented approaches in terms of

impact on the overall delay of the VHO preparation phase under various conditions. Lastly, Chapter 9 concludes the dissertation and provides insights for future work.

2. Context-Aware Connectivity Management in Light of ANM

2.1 Basic Concepts considering Context-Awareness, Cognition and Autonomicity

Context is any information that assists in determining any situation(s) related to a user, network or device [**35**] and can be distinguished to static and dynamic context and the levels of abstraction, as it is presented with more details in the following. Dynamic context in comparison to static context is more time-variant and thus more difficult to predict. Static context may include user's application preferences, the list of RANs serving a given area, security policies and cost. Dynamic context may include resources availability context (depending on the current network loading conditions), user location, MN velocity, battery power etc.

In the heterogeneous and complex 5G networking ecosystem, users should be able to have contextualized, proactive and personalized access to services everywhere, promoting the Quality of Experience (QoE). The term *context awareness* refers, in general, to the ability of computing systems to acquire and reason about the contextual information in order to be able to adapt the corresponding applications accordingly [**25**]. Hull et al. [**36**] described context awareness as "the ability of computing devices to detect, sense, interpret and respond to the aspects of a user's local environment and the computing devices".

Context-aware applications have the ability to adjust their behavior according to a different situation or condition without explicit user intervention, while situation awareness can be seen as the perception of an entity's situation to anticipate its needs/demands [**37**]. To achieve situation awareness all context conditions that describe what is happening should be known [**37**]. Context-aware systems have the ability to acquire and apply knowledge of context-

based information obtained by sensors [**37**]. We consider context-awareness as a fundamental autonomic feature, related to the Information Collection phase, as it is going to be discussed in detail, later on.

Cognition is related to intelligence and has been employed to enhance the effectiveness in network management solutions. The cognitive network concept is described in [**38**], as encompassing networks that can perceive current network conditions, plan, decide, act on those conditions, learn from the consequences of these actions and follow end-to-end goals. This feedback loop implements a learning model, in which past interactions with the environment guide current and future interactions, resulting in intelligence enhancements. Furthermore, in [**39**] it is claimed that cognition is mostly related to the inference plane, being driven by sensors, related to network planning and optimization and being differentiated from "involuntary functions" related to the management plane and configuration management, which is being driven by the "effectors". In other words, this second approach differentiates cognition from network management execution.

With respect to ANM, the ultimate aim is to create *self-managed networks* to overcome the rapidly growing complexity of networks. In 2001, IBM presented the autonomic computing framework, describing a system with 'self-x' properties, such as self-management, self-configuration, self-optimization, and self-protection [**40**]. Essential characteristics of an autonomic computing system include the capabilities to perceive its state and the state of its environment, to react accordingly to specific stimuli and to optimize its performance based on the reported status and stimuli. It is noted that autonomicity is frequently discussed by means of drawing analogies to biological entities, such as the human autonomic nervous system, so the relevant terminology can be metaphorically related to functional and/or structural aspects of a living organism [**41**], [**42**]. Correspondingly, the vision towards autonomic networking includes the following four closed control loops [**43**]: *sensing* (or monitoring) changes in the

network and its environment; *analyzing* changes to achieve the goals; *planning* reconfiguration if goals cannot be achieved; and *executing* those changes and *observing* the results. The operation of the control loops is enhanced by adding learning and reasoning processes, as well as by employing a well-structured knowledge base**.**

ANM enables the system to evolve and to adapt to changes, in terms of either business objectives or users' requirements. For this reason, ANM introduces rules to formalize the description of operations of various network elements in response to changes in the environment [**23**]. These rules are generally implemented by policies, defined (initially) by network administrators, guiding the behavior of network components. A typical advantage of policy-based network management systems is their ability to reconfigure and adapt their behavior by modifying the applied policies at runtime, without suspending system operation.

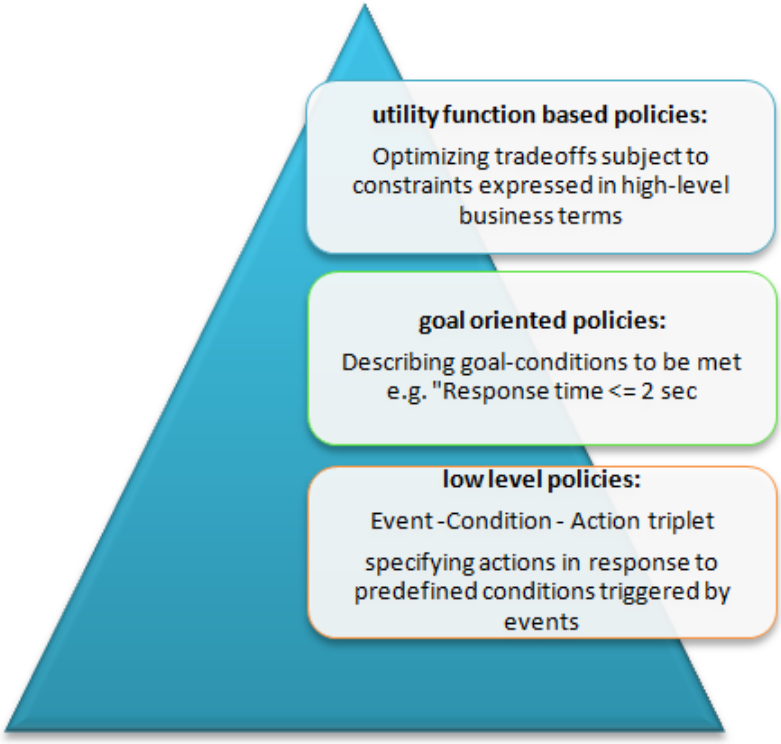

Figure 1. Policy-based network management hierarchy.

Policies at the lowest level are typically defined by the *Event-Condition-Action* triplet, which specifies the actions that have to be taken in response to predefined conditions, triggered by events. At the next level, goal policies are defined, which describe the goal-conditions that should be met, e.g. ‘‘Response time not greater than 2 sec” [**44**]. At an even higher level, some efforts have been dedicated to model system control using utility function based policies. Utility functions provide a natural and advantageous framework for achieving self-optimization in a dynamic, heterogeneous environment [**45**], [**46**]. Given a utility function, the system must use an appropriate optimization technique to determine the most valuable feasible state by tuning system parameters or reallocating resources, considering also aspects such as cost [**47**], [**48**], [**49**]. Additionally, utility functions allow degrees of flexibility in selecting different levels of QoS, matching the needs of different applications or user classes [**50**], [**51**].

By putting the various approaches just mentioned together, policies can be organized according to their purpose, forming a hierarchy, as depicted in Figure 1. Significant autonomic network management architectural frameworks are based on policies, such as Autonomia [**52**], DRAMA [**53**], Unity [**54**], ACCORD [**55**], CA-MANET [**56**], AutoI [**57**], ANA [**58**], and FOCALE [**43**].

In recapitulation, it can be stated that the concepts of cognitive networks and autonomic network management respond to almost the same expectations. Two differentiating factors on these definitions may be considered: the extent to which intelligence can be considered an axiomatic property for autonomicity; and the consideration of including network management execution among the cognitive networking tasks [**21**]. Based on IBM’s definition for autonomic computing, the basic self-x properties include awareness, adaptivity, proactivity and optimization, thus it can be argued that a system does not have to be intelligent to implement autonomic features, although intelligence can advance its overall degree of

autonomicity. On the other hand, in recent research papers concerning autonomic network management, learning and intelligence are being considered as fundamental dimensions of autonomic systems [**23**], [**22**], [**24**], [**8**]. In our point of view, considering a holistic autonomic approach, cognitive functions shall be considered as part of the autonomic network management framework of a FI system.

Lastly, a similar concept to ANM is the Self-Organizing Networks (SONs), which are able to independently decide when or how to trigger certain actions based on continuous interaction with the environment [**26**]. However, SONs do not currently include proactivity [**26**], which is considered as a fundamental feature of ANM, as it is presented in Chapter 3.

2.2 Taxonomy and Classification Framework for VHO Management in View of ANM

VHO management in the era of 5G concerns user mobility among multiple radio access technologies, multi-layer and even multi-operator dense network scenarios, where the user may have to perform multiple vertical handovers during the connection-time to switch among different cellular layers (e.g. macro-small cell) and/or radio interfaces (e.g. 4G, 5G, WiFi). Therefore, VHO management is a considerably more complex process than the management of horizontal handovers enabling user mobility in a single radio access network, as due to this high degree of heterogeneity, interoperability issues are posed [**6**]. In Figure 2 and in Figure 3, the different HO types are depicted existing in a HetNet environment, as well as, the effect on complexity according to different handover types. Also, there are other aspects contributing to the increased complexity, including the need for accommodating application demands and user preferences and for exploiting the capability of handling multiple active network interfaces concurrently.

One way to address these challenges is by introducing context aware MNs that self-manage their mobility patterns, towards meeting QoS requirements and maximizing user satisfaction.

In particular, the support of connectivity management between macrocells and femtocells dictates migration from network-controlled to autonomous, self- and environment-aware MNs, which can be founded on the use of cooperative and cognitive radio strategies [**59**]. Such functionality may assist in the neighbor cell list discovery and the cell reselection. For such operations the serving cell configures the MN to perform signal quality measurements to acquire the system information of the new cell [**59**]. In general, the context-aware MNs just mentioned, may be assisted by further components of the handover management architecture, higher up in the network hierarchy that provide/enforce appropriate policies to the MNs, in order to achieve global optimization goals (e.g., load balancing).

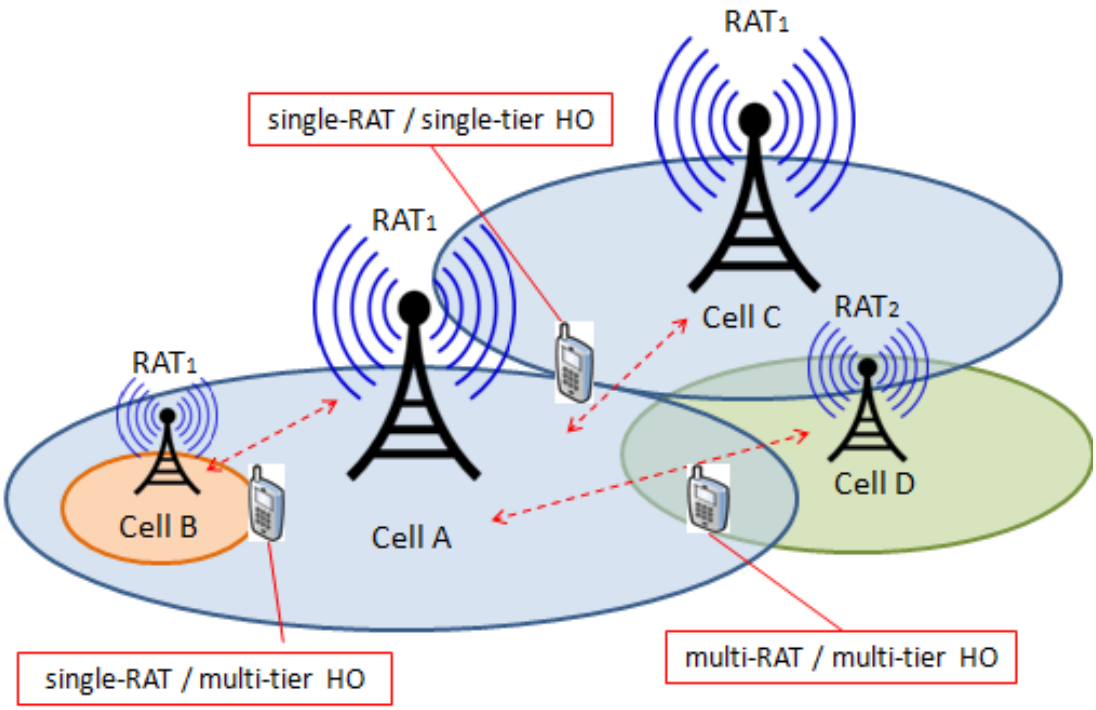

Figure 2. Different handover types in a HetNet environment.

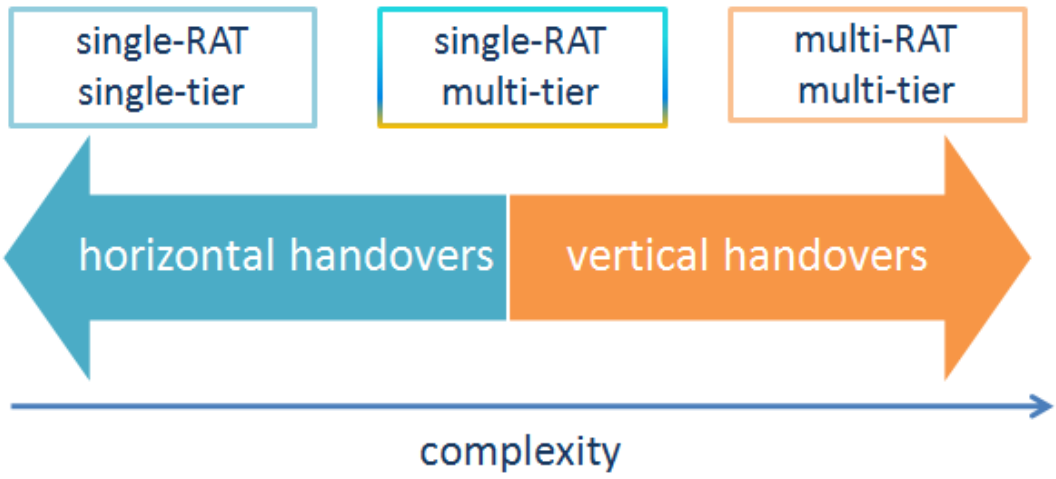

Figure 3. Handover types versus complexity.

It is noted that current research directions considering the increasingly denser and unplanned network layout, promote context-aware strategies that are not necessarily autonomic. For example, [**60**] proposes a strategy to minimize unnecessary handovers, aimed at multi-tier cellular networks, combining both user-location awareness and cell-size awareness, while [**61**] develops a velocity-aware solution via stochastic geometry, which resolves handover rate problem in dense cellular networks. Further optimizations can be achieved by splitting the control plane and user plane, using phantom cells. This has been proposed as a potential solution to minimize network control overhead in 5G networks [**62**]. Such solutions may be enhanced by incorporating elements from the autonomic networks that can enable distributed context awareness, processing and decision making.

In line with the trends just discussed, this section discusses handover management frameworks in light of ANM, assuming context aware MNs self-managing their mobility behavior (at least to a degree) according to policy-based management principles. As part of this discussion, we introduce a new taxonomy of the relevant architectural components.

In principle, the structure of the media independent handover mechanisms can be taken as a basis for organizing the discussion of relevant autonomic management features. With respect

to this structure, and according to established VHO frameworks (such as the IEEE 802.21 [**12**] and 3GPP Access Network Discovery and Selection Function (ANDSF) [**63**]), the handover management procedure can be separated into three phases: *handover initiation*, which contains network discovery, network selection and handover negotiation, followed by *handover preparation* which contains layer 2 connectivity and IP connectivity, and then complemented by *handover execution*, which includes handover signaling, context transfer and packet reception. However, here we organize the relevant operations in a slightly different manner that enables us to highlight the autonomic character/elements of the handover management procedure. A similar organization has been followed in [**27**].

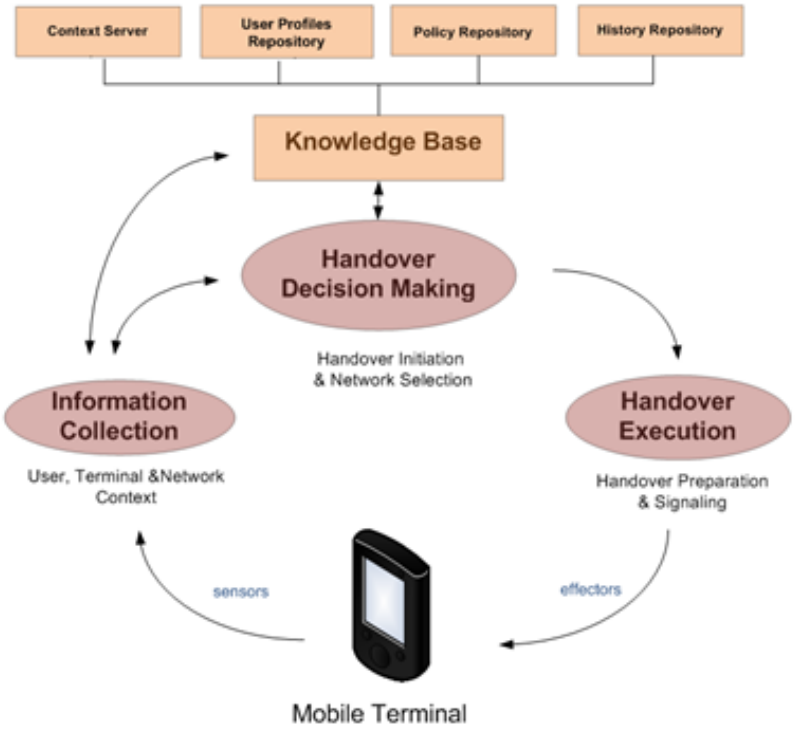

Figure 4. Phases of the autonomic VHO management and associated architectural components.

Specifically, the operations are grouped into the phases of *information collection*, being linked to a *knowledge base*, followed by the *handover decision making*, which includes

handover initiation and network selection processes and its corresponding algorithms, itself followed by *handover execution*[‡], which includes handover preparation and signaling. In the context of these redefined phases, the handover management complies with the autonomic management principles of monitor (information collection), analyze & plan (handover decision making), and execute (handover execution) functions.

Figure 4 illustrates the interaction of these phases and their connection with key architectural components. The alternative grouping of phases just presented reflects better the autonomic control loops involved, while remaining fully aligned with the aforementioned, more conventional, grouping (i.e. handover initiation, preparation and execution), in the sense that all individual actions are included in both groupings. As a summary, Figure 5 depicts relevant attributes discussed in the following.

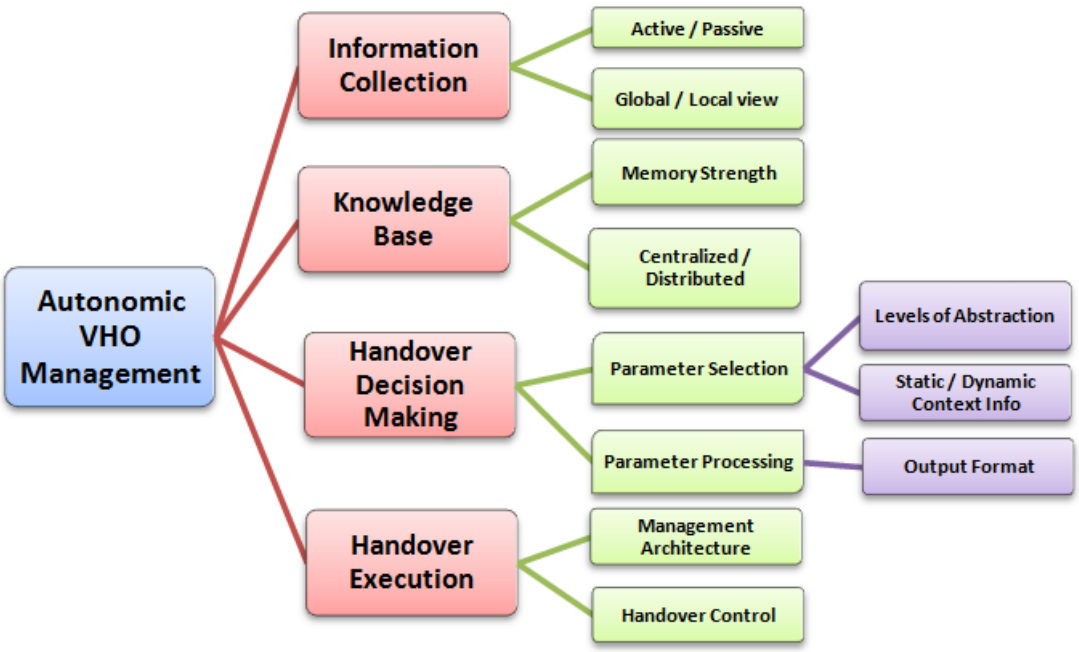

Figure 5. Key attributes/properties of the autonomic VHO management phase

‡ The term 'handover execution' here refers to the *Monitor-Analyze-Plan-Execute* loop in ANM, instead of the actual *handover execution phase* within the standards' based handover management procedure mentioned earlier.

2.2.1 Information Collection

The information collection process gathers the required user, terminal and network context, in order to provide to the MN self- and environment-awareness. This process is critical, as it constantly provides appropriate information to the 'analyze & plan' functions of the handover decision making phase, indicating the need for a handover initiation and assisting in the network selection. According to [**28**] and [**64**], contextual information encompasses user context, including user-related information such as preferences, priorities and profiles history. Another type of context is terminal-related information, such as power status, physical mobility parameters (e.g., distance, location), Received Signal Strength (RSS) and Signal to Noise plus Interference Ratio (SINR) measurements, as well as information related to running applications (e.g., QoS requirements).

Furthermore, network context may be included, providing indicators of the quality and the availability of resources of neighboring networks (through metrics such as bandwidth or throughput), or provider context (e.g., cost, security management, etc). Finally, another important type of context relates to handover performance, including parameters such as handoff latency, decision latency, execution latency, degradation rate, and improvement rate.

Concerning autonomic handover management solutions, user-related context plays a significant role, permitting the maximization of the user satisfaction by taking into account the user preferences. More precisely, the handover decision making module uses the collected information to evaluate the available access networks and to select the most capable network, satisfying at the same time the user's request at a particular time (e.g., "maximize throughput, but also minimize monetary cost"), referred as *Always-Best-Satisfying* (ABS) network [**65**], [**43**].

This more elaborate consideration of user preferences refines the concept of an *Always Best Connected* (ABC) device. Therefore, it is important for the information collection module to maintain user profiles, in order to be able to accumulate the user-related context.

It is noted that the volume of collected data must be post processed and/or converted in a form suitable for later use. In particular, raw measurement data should be converted to a common format, understandable by subsequent decision making processes. Also, to avoid overloading the information collection and knowledge base components with raw measurement data, *filtering* is needed [**66**]. Consequently, there is a trade-off between precision and measurement load [**67**].

In autonomic handover management, information collection may be characterized as *active* or *passive*. In the active case, the MN itself can initiate data collection periodically, including the issuance of testing messages. By contrast, in passive information collection status capturing is initiated and (more generally) coordinated by components at the network side [**68**].

In order to make the information collection process even more effective, *global statistics* of network-wide scope could be collected and analyzed, about neighboring MNs and their experiences with the different access networks available in the area [**32**], [**8**]. The global statistics gathering/analysis may potentially employ cloud computing services and/or big data analysis [**69**]. Since the network-wide view is built by sampling local views from various MNs within the network, a particular MN can utilize the global view to compare against its own status and to potentially self-adjust. For example, the global knowledge could provide hints to MNs for generating dynamically optimized policy parameters. More generally, information about neighboring MNs can be exploited in the decision-making process (e.g., to determine the right time to initiate a VHO) and to identify the best course of action with respect to, e.g., QoS or energy efficiency.

In the sense just mentioned, global statistics lead to an advanced awareness that could promote the adaptation and learning processes, leading to optimal handover decisions and ultimately improving the QoS for the end users. However, it is noted that collecting and maintaining network-wide global statistics may require more computational and memory resources and may lead to increased power consumption and signaling overhead. Therefore, the information collection should strive for balancing the trade-off between the extra overhead and the more comprehensive network view.

2.2.2 Knowledge Base

The knowledge base stores user, terminal and network context received from the information collection module, making this information available and accessible to other autonomic handover management entities that require it, such as the handover decision making functions (as shown in Fig. 2), contributing to the cognition loop. Based on [**67**], we classify the components of the knowledge base, into four logical groups, depicted at the top of Fig. 2 and further discussed in the following.

To begin with, the *Context Server* stores current terminal, application and network context, logically divided in two parts: the *Service Information Base* and the *Resource Information Base*. The first part contains information about the service instances activated by customers, such as parties involved (customers and service providers), rules regulating the service delivery, types of resources needed, amount of each resource type needed in each occasion, billing plan for the service, and operation history. The second part maintains an up-to-date account of the type and quantity of currently available resources.

The *User Profiles Repository* contains information related to the users of the mobile device, such as user preferences, user history, list of the subscribed services, and updated billing information. It is noted that, while the User Profiles Repository could in principle be

regarded as a part of the Context Server, we keep it separate, in order to emphasize its importance in autonomic handover management.

The third logical group of the knowledge base, namely the *Policy Repository*, contains information related to policies, for use in the handover decision making [**63**]. Complementarily, the *History Repository* logs information about previous handover decisions, such as parameters employed, cause that triggered the handover, time of occurrence, parties involved, target network selection, and effect of this selection. Using this log, the current situation may be correlated with previous comparable ones, so that decisions can be made faster, saving time and computational power.

Depending on the implementation, a knowledge base may be classified as *centralized*, when the entire knowledge base is a single central entity residing at the network side, or *distributed*, when the knowledge resides at various places, mostly at the edge of the network, or even at individual MNs. The cloud computing concept is conformal with the centralized knowledge base paradigm, offering centralized data storage and processing through remotely deployed server farms and software networks [**70**]. However, the traditional centralized cloud computing architecture may fall short in meeting the strict latency requirements for mobility management in a 5G network environment. Edge, mobile edge (i.e., Mobile Edge Cloud (MEC)), mobile cloud and fog computing concepts, which use computing resources and storage at the edge of a network [**71**], can be used as alternatives related to the distributed knowledge base paradigm, potentially offering a higher delay efficiency. Such distributed knowledge base paradigms could facilitate the MNs to store individually essential information about their mobility for later use [**27**], further promoting the concept of self-management.

Furthermore, another attribute of the knowledge base, related to the history repository, is *memory strength* [**8**], referring to the ability of the system to remember past behaviors,

significant events, corresponding reactions and results, towards assisting the system in its current and future management decisions.

2.2.3 Handover Decision Making

2.2.3.1 Parameter Selection

The handover decision making can be considered as the core phase of the VHO, since it is in charge of analyzing the context collected by the information collection phase and planning the actions to determine the best handover target [**28**]. This phase includes handover initiation and network selection processes and the corresponding algorithms. In the autonomic context of interest here, the handover decision making also includes cognitive self-learning mechanisms that enable the system to meet the forthcoming needs, promoting self-optimization and self-healing. In reflection of this fact, the handover decision making phase can be organized into two distinct steps: the parameter selection and the parameter processing.

The *parameter selection* exploits the context gathered in the information collection phase, towards selecting suitable parameters from a given set/pool (determined by the user or by a policy in effect). The selected parameters are fed as input to the parameter processing algorithms, essentially determining the criteria for the decision making therein. The versatility of the parameter selection is characterized by two attributes: context time variability and levels of abstraction [**32**]. The context time variability expresses the potential for including in the selected parameters set both static and dynamic context. The levels of abstraction refer to the capability of the autonomic system of jointly treating multi-layer context in uniform, abstract terms and thus the capability to make parameter selections spanning several layers among the physical, link, network, transport and application ones [**28**], [**72**].

2.2.3.2 Parameter Processing

Since, as already mentioned, the *parameter processing* receives parameters selected on the basis of gathered context, the decision making therein becomes context aware. In particular, the context encapsulated in the selected parameters drives the parameter processing algorithms, towards making optimal decisions (with respect to multiple criteria). With respect to the algorithms themselves, and considering the current state-of-the-art of context-aware VHO decision making solutions, parameter processing methods can be classified into the following four distinct approaches: a) the decision function (DF) approach (including simple DFs and Multiple Attribute Decision strategies (MAD)); b) the Markov decision process (MDP) approach; c) the policy-based (PB) approach (including Finite State Automata (FSA)); and d) approaches based on fuzzy logic (FL) or neural networks (NN). Each of these approaches is discussed further in the following.

DF strategies use the selected parameters to calculate the values of specific decision functions that assess the merit of individual alternative actions. The decision simply selects the action with optimal merit. In this sense, DFs can be regarded also as award, cost or objective functions. For specific related applications of the concept, see [**73**], [**74**], [**75**], [**76**]. The prime advantage of this approach is simplicity. In particular, for cases involving only a small number of parameters, network selection may employ a simple DF evaluating the weighted sum of values derived from the selected parameters (repeatedly, for each network in the service area of a user).

A more sophisticated distinct sub-family of decision function-based methods involves MAD strategies. These combine and evaluate multiple decision criteria simultaneously, dealing efficiently with complex problems, and providing high flexibility [**27**], [**77**]. MAD strategies can be classified into several groups, including:

- the Simple Additive Weighting (SAW) [**78**], involving a larger number of parameters than the simple DFs, where the score of a particular network is determined by the weighted sum of all the attribute values;
- the Techniques for Order Preferences by Similarity to Ideal Solution (TOPSIS) [**78**], where the preferred network is the one closest to the ideal solution and farthest from the worst case solution;
- the Grey Relational Analysis (GRA) [**79**], which ranks the candidate networks and selects the one with the highest ranking; and
- the Analytic Hierarchy Process (AHP) [**80**], which decomposes the network selection problem into several sub-problems and assigns a weight value for each sub-problem.

According to [**81**], the advantage of AHP solutions is their strong robustness for solving problems with complex hierarchical structure. On the other hand, considering problems with relatively simple hierarchy, SAW is less complex and thus preferred. In [**82**], AHP is used to determine weights to the selected parameters (bandwidth, delay, jitter, and Bit Error Rate (BER)), applied to several MAD algorithms, including SAW, TOPSIS, and GRA, while a performance comparison between them is performed. Results show that SAW, and TOPSIS provide similar performance for conversational streaming and interactive traffic classes, whereas GRA provides a slightly higher bandwidth and lower delay for the interactive traffic class.

Another parameter processing approach involves MDPs. The handover problem under consideration is formulated with the objective of determining the action that maximizes the total expected reward per connection [**83**], [**84**].

To this end, Deterministic Markovian (DM) decision rules are employed. These are functions that specify the action choice when the system occupies a particular state at a specified decision epoch. Transitions from state to state are governed by a Markov chain,

which captures memory effects. The state information includes the current network status plus availability of other networks in the area. The time between transitions corresponds to the time between successive decisions. To specify the MDP, one should calculate the probability of transition from one state to another. The transition probabilities can be estimated by the network operator based on gathered statistics.

Several particular applications have been based on this general framework. For example, in [**85**] the transition probabilities are assumed to depend on the suitability (rank) of candidate networks in relation to each decision parameter and on the weight of each such parameter. Analysis of this model enables the determination of the optimal candidate network [**85**] under a particular set of state conditions, while the derived results can be exploited for future decisions.

In [**86**], the optimal decision rules are constructed by means of AHP, combining the benefits of MDP and MAD approaches. In [**83**] the calculation of the optimal decision is performed by the operator offline and is periodically updated whenever spare processing capacity is available at the network access controller.

In general, the update frequency of the Markov chain transition matrix and the flexibility and adaptability of the decision parameters are crucial factors determining the suitability of MDP for use in autonomic handover management. Apart from the core handover management functions, however, MDP techniques may also be used for user (physical) mobility modeling with a Markov chain, towards extracting the user's mobility patterns from a historical mobility trace [**86**]. In this way the next possible location of a user can be estimated, which will determine the next possible network connection(s), optimizing handover performance.

The third approach to parameter processing involves policy-based decisions. In this case, network selection proceeds by determining the most suitable network according to a specific

set of policies. For policy conflict resolution, FSA can be employed, where policies can be represented as deterministic transducers [**66**], used to resolve potential conflicts, both static and dynamic, among the different policy rules. At a next step, a decision function (Tautness Function (TF) [**87**]) is formed, to indicate how tautly a condition fits to an event. Subsequently, priorities are assigned to the conditions, depending on their probability to occur.

The common drawback of all three parameter processing approaches already reviewed is their inefficiency to handle a decision problem that involves ambiguous decision criteria. To remedy this deficiency, in specific scenarios FL or NN could be used as an intermediate step. FL-based strategies convert parameters into fuzzy sets.

A set of fuzzy rules are applied utilizing a series of branches roughly analogous to ordinary IF-THEN clauses, producing a decision set (growing or shrinking as successive rules are applied) that is subsequently mapped into a single-valued quantity. Related applications can be found in [**27**], [**65**], [**88**], [**89**], [**90**].

On the other hand, NN are usually employed with only one parameter and one type of handover policy (i.e., "keep WLAN connection when it is available"). However, NN architectures require training delay and prior knowledge of the radio environment [**27**]. Related applications can be found in [**91**], [**92**].

Finally, another important aspect of parameter processing relates to the *output format* of the network selection, indicating the target network candidate(s). For example, the output may be a list of the candidate networks in prioritized order, where the top of the list represents the one with the highest significance/weighting factor according to the predefined criteria [**27**].

Alternatively, the output format could specify only one candidate network, selected by a policy-based framework [**66**]. Moreover, when multiple active interfaces are supported, the output could specify the appropriate network interface for each application [**65**].

2.2.4 Handover Execution

2.2.4.1 Handover Control Method

The handover execution process implements the VHO management and control [**28**]. In this study, we are concerned with the control methods and management architectures considering state-of-the-art autonomic VHO management solutions, assuming these are distributed enough to enable the MN to make (at least some) decisions on its own, promoting self-management. Accordingly, fully centralized and/or fully network controlled management approaches are not considered in the following, being out of scope.

In particular, self-management is regarded as an essential property of autonomic VHO management, giving to the MN the ability to control its own context and enabling it to determine the appropriate time to execute handovers [**93**]. Furthermore, self-management promotes adaptivity, flexibility and self-optimization to the decisions of the MN [**8**]. According to the most recent trends, distributed handover management may provide a paradigm most congruous to the need for handling effectively the complexity of the FI environment and the emerging 5G networks, avoiding at the same time a single point of failure (characteristic of the classical centralized approaches, frequently together with high latencies and signaling overhead) [**69**], [**94**].

In general, the autonomic handover process may be characterized by the entity that is responsible of controlling it. It is characterized as *mobile controlled* [**28**], [**27**], [**95**] when the VHO initiation and decision is fully controlled by the mobile device. This is a flexible solution that enhances user satisfaction. The disadvantages are that the MN must possess advanced computational capabilities, which also lead to increased power consumption. Alternatively, the VHO may be characterized as *network assisted*, if the handover initiation is done by the mobile device, but the network selection, or a part of it, is implemented by the

network, making use of the information services and undertaking the heavy programming tasks [**66**], [**65**]. Finally, the VHO is characterized as *mobile assisted* when it is initiated by the network, but assisted by the mobile device [**96**].

Beyond the control-related characterization just discussed, the structure of the management architecture is important, since it affects the scalability, performance, intelligence and overall autonomicity of the system [**8**], [**68**]. This structure can be classified into three basic categories: flat, hierarchical and hybrid.

2.2.4.2 Management Architecture

The *flat* approach refers to fully self-managed MNs, where autonomic handover managers (AHMs) are assumed to reside only in the intelligent MNs (as depicted in Figure 6). This distributed type of management architecture addresses the limitations of centralized management with respect to fault-tolerance and scalability, advancing autonomicity. However, this approach raises challenges in the domain of distributed information management, system-wide coordination, security, and resource provider's policy heterogeneity. It may also put on MNs excessive requirements in terms of computational capabilities and power consumption. Examples of flat autonomic handover management approaches are found in [**27**], [**95**].

In the *hierarchical category***,** a main AHM supervises a set of multiple lower-level AHMs. Thus, a coordination management overlay should be defined, to arrange the operation of lower-level autonomic managers. Considering hierarchical architectures, intelligence resides in both the terminal and network sides, avoiding excessive complexity at the MN. In general, hierarchical architectures can be *centralized hierarchical* or *distributed hierarchical* [**8**], depending on whether the main AHM resides in the network side or in the MN side. Hierarchical autonomic architectures consider a distributed manager level (see Figure 4) [**8**]. A significant advantage of using distributed hierarchical architecture is that MNs present a

higher degree of self-management and thus they can support autonomic handovers more efficiently. Also, more personalized management policies can be deployed at the autonomic manager of each MN. Examples of such approaches are [**66**], [**73**] and [**65**].

More specifically, according to [**65**] and [**73**], the main AHM resides in the MN. However, some functionalities are placed also to the operations and support system (OSS), where network monitoring is performed. Furthermore, a context server that resides in the core network collects the relevant contextual information from the various repositories and assists in the handover decision, in response to requests from the MN. Similarly, according to [**66**], those components that involve operator's management or high computational cost are located in the core network to minimize the complexity of the MN. Such tasks include policy definition, storage, and conflict resolution. These network-side components assist in handover decision, always with the coordination of the MN.

In addition to the two categories already discussed, there are various hybrid architectures combining the previously mentioned concepts to a varying degree. In autonomic *hybrid* architectures, some self-organization and self-optimization algorithms, mostly those related to tasks with local scope, are running locally on the MN, while the tasks with wider scope (global network view) are being managed by a central managing authority on the network side (usually at the base station or in a cluster of base stations, as shown in Figure 6).

It is noted that while the distribution of functionality just discussed is at present considered to be fixed, future autonomic VHO management architectures may have dynamically adjusted structure, towards an increased potential for customization [**24**]. In general, hybrid architectures achieve load balancing and traffic management, hiding the complexity from the MN.

Examples of hybrid architecture can be found in [**97**], where autonomic MNs are assumed to cooperate with autonomic base stations and access points in order to make optimal

handover decisions and QoS-aware resource management. Also, [**93**] proposes a scheme for vertical handover decision making that leverages the cooperation between the MNs and a controller, which manages a cluster of different access networks locally available. This controller is also responsible for resource control and load balancing among the MNs.

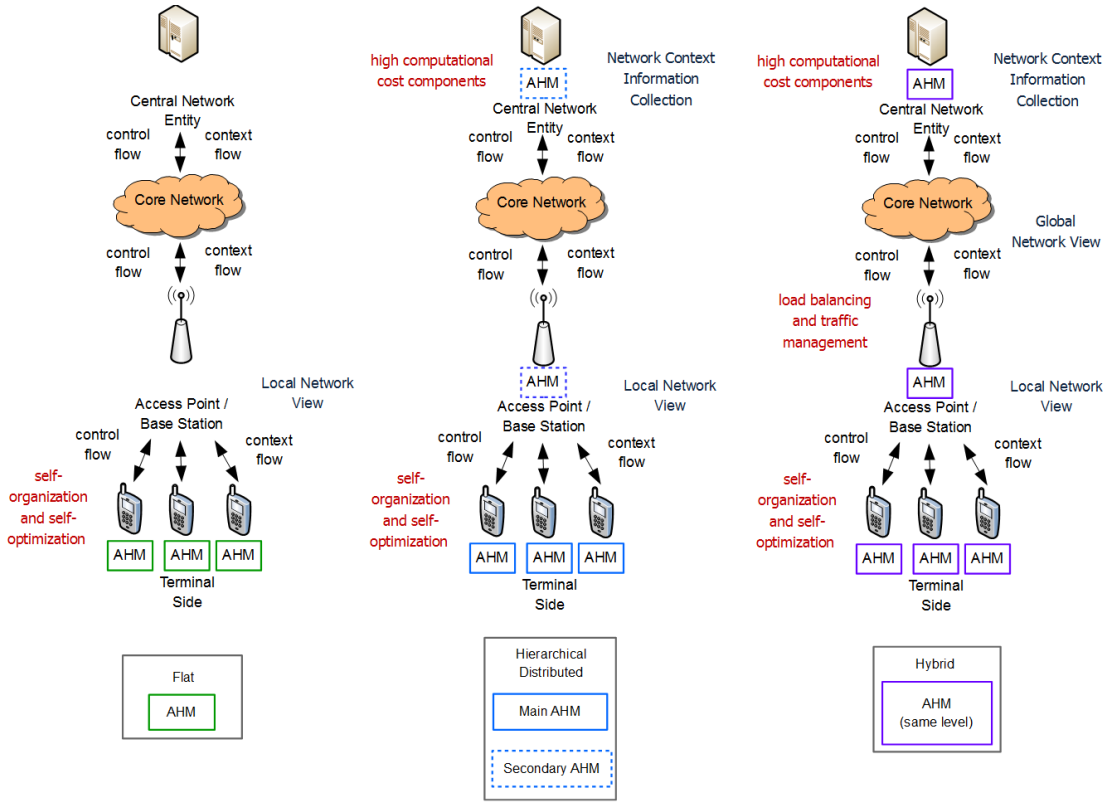

Figure 6. General architectural model for VHOs with multiple levels of AHMs.

3. Autonomic VHO Features Towards Self-Optimization And Robustness Issues

3.1 Autonomic VHO features towards overall self-optimization

We now turn to a number of important features, which characterize autonomic handover management and jointly lead to performance optimization. Firstly, we discuss the nature and effect of each of these features, in correlation with the taxonomy of Chapter 2. Secondly, we deal with the issue of robustness, also discussing how the autonomic features may be exploited towards more robust handover decision making.

Autonomic VHO management aims at performance optimization, related to seamless mobility and user satisfaction. The deployment of autonomic features to automatically manage, optimize, and adapt the management of operations can significantly improve the resulting performance [**47**], [**98**]. More specifically, the combination of awareness, adaptivity, flexibility and proactivity drive the system to performance improvements and enable the system to select the best choice among a set of available alternatives, advancing the system's overall self-optimization, which can be described as the objective of autonomicity. The functionality of individual autonomic feature and the inter-relations between them towards the optimization of VHO management performance are further described in the following. A summary is depicted in Figure 7.

3.1.1 Awareness

This is a fundamental property, present in most autonomic functionalities (i.e. see context-awareness in Chapter 2.1). Awareness is primarily related to the monitor function of the information collection phase and the associated knowledge base and is most directly exploited in the parameter selection step of handover decision making. The term refers to self

and environment awareness, addressing information collection from the MN, the networks and the user [**47**]. Awareness is expected to trigger a `prompt reaction' associated with the handover execution, thus closing the autonomic loop in Figure 4.

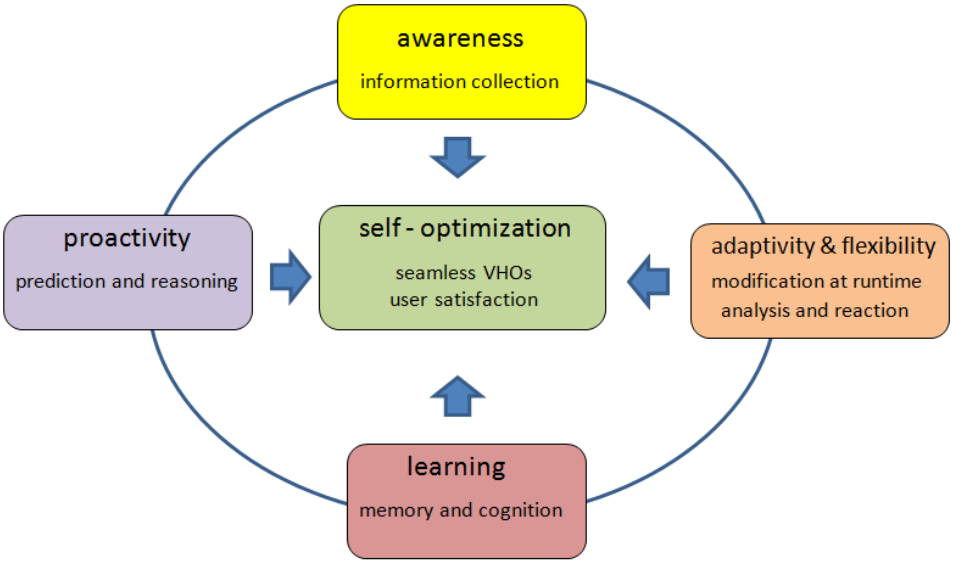

Figure 7. Autonomic VHO management features towards overall self-optimization.

Specifically in connection with VHO management, an enhanced level of awareness is positively correlated with the ability of the system to extract contextual information from multiple layers (a notion linked to the aforementioned levels of abstraction), enabling the consideration of the QoS requirements of running applications. The level of awareness is also related to the frequency of parameters monitoring, which affects the precision of the selected parameters used in the VHO decision making. Finally, awareness also affects the capability of the system to support an adjustable monitoring process. This is further discussed in the following, in connection with adaptivity & flexibility.

As a concrete example, [**65**] demonstrated that a high degree of awareness resulted in enhanced (by more than 20%), end user satisfaction metrics when compared against other algorithms not considering user preferences.

3.1.2 Adaptivity & Flexibility

In the more general context of autonomic network management, adaptivity deals with the ability of the network to analyze changes indicated by current events (perceived due to the system's awareness) and to decide why, when, where and how a reaction should take place [**8**]. Thus, adaptivity involves the 'analyze' and 'plan' components of the autonomic loop. For example, adaptivity may trigger changes to the frequency of measurements during the information collection phase, and may promote adjustments to the parameter selection and parameter processing methods, according to environment changes and system needs [**22**], [**47**]. Towards this direction it is noted that adaptivity could be further enhanced by the use of biologically inspired solutions. For example, swarm intelligence has been employed in autonomic network management, addressing load balancing and route construction and maintenance [**22**].

The achievable degree of adaptivity depends on the level of flexibility [**32**] (characterized as *limited* or *advanced*), which is related to the capability of modifying at runtime parameter selection and processing methods for use in handover decision making. The term 'advanced flexibility' refers to the capability of dynamically adjusting the set of said parameters, potentially including newly identified parameters at run-time, without requiring modifications to the implementation of either the support system or the application logic. The term 'limited flexibility' characterizes approaches that are narrower in scope and involve a predefined parameters' domain, determined during system design.

Obviously, VHO solutions featuring advanced flexibility equip the system with a greater ability to evolve, and thus improve awareness and adaptivity, making the system capable to adjust to a changing environment [**99**], [**68**], [**8**]. For example, the adaptive approach used in [**66**], reduced handover latency, resulting in close to seamless connectivity on the move.

3.1.3 Learning

The learning functionality is part of cognition and equips the system with the ability to remember past behaviors, or problems and their solutions. This ability, in turn, helps the system to gain experience (to an extent determined by the memory strength of the knowledge base) that may be utilized in the decision making, in combination with adaptivity. The knowledge of behavioral trends and occurrence patterns of conditions/scenarios is valuable, especially in highly dynamic environments, as it can dramatically enhance the system performance, by identifying frequently repetitive patterns of actions and behaviors. Towards this end, artificial intelligence techniques may be employed, such as neural networks [**23**]. It has been shown in [**95**] that the exploitation of historically available information led to an improvement of about 50% in the mobile handset's battery autonomy and to about 25% lower content downloading times and network usage costs.

3.1.4 Proactivity

Proactivity signifies the use of preventive measures to maintain a target level of system performance (by means of an appropriately and timely initiated handover procedure), based on the analysis of the current state and on the anticipation of events and their effect on the system. Anticipation is a cornerstone of proactive computing, promoting actions in the direction of *future prediction*. Proactivity involves the 'analyze' and 'plan' components of the autonomic loop, utilizing data from the information collection process, which is equipped with awareness. Proactive techniques focus on *context aware operation, statistical reasoning, and intelligent data-handling* [**100**]. By proactively collecting and analyzing predicted information about e.g., the link status or the battery status, the resulting VHO decisions can be optimized [**29**], [**101**].

Therefore, proactive systems exploit context for responding faster and more efficiently to specific stimuli, providing further benefit if used in conjunction with learning techniques.

Statistical reasoning techniques such as Hidden Markov Models, genetic algorithms, and Bayesian techniques, can be used instead of traditional deterministic methods. For example, [**86**] computed the user's mobility regularity from the historical trace of the user using an MDP process, toward providing estimations for the next possible location of the user, subsequently exploited for making more robust VHO decisions. Evaluation results in [**86**] showed that this proactive strategy, used in conjunction with a multi-attribute decision algorithm, achieved around 50% better performance gains (in terms of throughput and latency) compared to a baseline greedy strategy. Other proactive user location estimation algorithms [**60**] and [**61**], resulted in minimization of unnecessary handovers, providing throughput gains up to 47% and 70%, respectively. In general, proactive features in autonomic network management promote network and resource availability, service level agreement compliance, and enhance user satisfaction.

3.2 Robustness Issues towards Stable and Efficient Decisions

As already mentioned, VHO management in a FI environment must cope with the heterogeneous, diverse and dynamic character of the target setting and the need to jointly consider many different sources of context. In such an environment, robustness (generally defined as the ability to achieve stable and efficient decisions [**102**]) becomes an important attribute of the VHO decision making process. The following subordinate paragraphs identify a number of robustness-related issues and review mechanisms to overcome them. Subsequently, we discuss how the autonomic features can contribute towards enhancing robustness. A synopsis of the relevant discussion appears in Table III.

3.2.1 Diversity of Parameters (Context Diversity)

The joint consideration of multiple sources of context creates the need for dealing concurrently with a diverse set of parameters. This, in turn, requires a methodology to enforce

a uniform representation, so that different parameters-characteristics (naturally involving different units) are expressed through comparable values. The way to address this issue depends on the type of the method used for the parameter processing step of the handover decision making. For parameter processing using DF (including MAD) or MDP approaches, *conventional parameter normalization* (CPN) techniques [**103**] are appropriate, while FL-based parameter processing naturally resorts to techniques employing fuzzification. Note that PB approaches do not require a uniform representation methodology, as each parameter is processed individually, through a relevant policy.

3.2.1.1 Conventional Parameter Normalization Techniques

Accordingly, CPN techniques can be organized in two categories [**103**]. The first one employs *absolute normalization*, where each parameter's value is individually scaled between 0 and 1, with respect to a given minimum and maximum value [**103**]. Examples of multi-criteria applications, which incorporate scales that conform to absolute normalization, can be found in [**103**].

For example, consider a decision problem, where there are two relevant criteria (i.e. monetary cost and delay), rated for two available networks (i.e. WLAN and WiMax). Following the previous example, cost can be normalized with respect to a given minimum value of 0, and a maximum value of 1 $/min, while, delay can be normalized with respect to a given minimum value of 0, and a maximum value of 100 ms. In this case, WLAN is slightly ahead of WiMax (see Table I).

TABLE I. Example of Absolute Normalization

	Cost ($/min)	Delay (ms)	Normalized Cost (weight = 0.5)	Normalized Delay (weight = 0.5)	Overall Performance
WLAN	0.2	45	0.8	0.55	0.675
WiMax	0.5	25	0.5	0.75	0.625

The second category employs *relative normalization* [**104**], [**105**], where the scores corresponding to parameter values associated with different options (i.e., the various candidate networks scores) are summed up and scaled to 1. For example, if network's A delay is 45 ms and network's B is 25 ms, then the network's A normalized delay results to 0.36 and network's B to 0.64, accordingly, as the bigger normalized score corresponds to the better network.

TABLE II. Example of Relative Normalization

	Cost ($/min)	Delay (ms)	Normalized Cost (weight = 0.5)	Normalized Delay (weight = 0.5)	Overall Performance
WLAN	0.2	45	0.7	0.36	0.53
WiMax	0.5	25	0.3	0.64	0.47

This is a more complex process, as all parameters have to be rescaled whenever there is a change to any candidate network's score. However, under relative normalization methods, the final result is more distinctive [**104**], [**105**], as it is shown in the previous example. Also, AHP users have typically employed relative, rather than absolute normalization. In fact, the traditional AHP recommends that scores for options relative to each criterion, should be determined in exactly the same way that criteria weights are determined; and weights are always relatively normalized.

3.2.1.2 Fuzzy-logic based Parameter Processing Techniques

With FL-based parameter processing, conversion of absolute parameter values to relative ones comes as part of FL's inherent capability for handling a decision problem that involves ambiguous decision criteria [**106**], [**107**]. The approach of FL is comprised of four steps [**108**]. The first step is the fuzzification. For example, if the delay of a voice call is 25ms, through the membership function the delay is identified as low or high [**65**]. The second step is the rule evaluation, e.g., "if delay is low and jitter is low, then quality of voice call is high". The third step is the rule aggregation, where every result is aggregated into one fuzzy set for

each output variable. The last step is the defuzzification, where the fuzzy sets are converted into appropriate output values. For example the output values can vary between “strong accept” and “strong reject”, acquiring numerical values between 1 and 0, respectively.

A special form of the parameter representation issue emerges when considering the QoS requirements of different applications. The QoS parameters should be treated differently by each application, as each one has its own QoS constraints [**73**]. For example, the jitter-related requirements for a voice call differ from those of a streaming application. Therefore, a different treatment for the jitter scores should be used in each case [**65**]. When absolute normalization is used, the appropriate upper and lower values should be identified for each case. For relative normalization, application-specific thresholds for the relative scores are needed, to ensure that that the decision yields acceptable values for each criterion, for each application. Similarly, for parameter processing involving PB approaches, different policies should be specified for each application. Finally, if FL is used, different membership functions should be used for handling the same QoS criterion in connection to different applications [**65**], [**107**].

3.2.2 Diversity of Criteria/Rules

An effective handover decision making should be capable of jointly employing multiple criteria/rules, and assigning different importance to each of these criteria, towards optimized decisions tailored to the environment. Specifically, [**109**] demonstrated that that properly assigning importance to criteria has a direct impact on the handover failure probability.

FL-based decision making inherently lends itself to the joint consideration of multiple criteria, through the definition of parallel rules that may be applied simultaneously, to obtain the desirable outcome [**65**], [**110**]. For example, the fuzzy rule "If bit error rate is low AND burst error rate is low AND packet loss ratio is low, then quality is Strong Accept" [**65**], ensures that a candidate network would be strongly preferred as the handover target when all

three conditions are satisfied. PB approaches can also handle groups of parameters according to different criteria, through relevant policies. However, in this case, the occurring conflicts have to be resolved.

For other parameter processing methods, employing CPN techniques, a different importance can be defined for each individual criterion [**109**], while, AHP [**80**] may be used in order to assign a different level of importance to each group of criteria. Indeed, AHP decomposes the decision problem into several sub-problems, making use of hierarchy. Different groups of criteria may be associated with different AHP sub-problems and their relative importance may be tuned through the assignment of corresponding weights [**107**]. For example, according to AHP, the first tier of parameters could include cost and QoS, associated with respective weights. The QoS could be further analyzed into a set of second-tier parameters, such as bandwidth, delay and jitter. The weights for the parameters in the second tier could be adapted according to the demands of each application and to user preferences.

It is worth mentioning that a number of VHO management proposals use initially FL followed by CPN techniques to employ AHP, (such as [**65**], [**107**]), combining the benefits of both approaches.

3.2.3 Context Uncertainties & Incompleteness

Another set of robustness-related challenges arises in connection with the ability of the decision system to cope with uncertainties and incomplete information. Uncertainties refer to the imprecise knowledge of context, particularly when successive measurements for the values of some parameters fluctuate beyond a level of tolerance. Incompleteness is associated with missing information, including the lack of information due to failures encountered during the information collection phase of the VHO management process.

One way to rectify the effects of uncertainties, particularly considering performance-related measurements, is by verifying the measured data against related data referring to other layers [**72**], [**111**]. This general concept is consistent with all methods for the uniform representation of parameters, including CPN and the methods appropriate for PB- or FL-based parameter processing algorithms.

The way to address incompleteness varies slightly, depending on the uniform representation method in use. For CPN or PB-relevant methods, an "average" value may be substituted for the missing one. For FL-based parameter processing, substituting a "neutral" value is the suitable course of action [**110**]. These general principles for handling uncertainties and incompleteness can be further enhanced by making use of the memory strength available at the VHO framework and of any available learning techniques, towards exploiting historically available relevant data.

3.2.4 Marginal / Borderline Cases

Robustness is important also for coping with cases where there are marginal differences among candidate networks that may lead to unnecessary VHO decisions. This phenomenon is frequently described as the 'ping-pong' effect [**27**], referring to repeated successive VHOs between the same two networks, which eventually leads to QoS degradation. A related phenomenon is the 'corner effect' problem [**112**], where the MN cannot assess correctly if a neighboring network is a suitable VHO candidate, due to poor line-of-sight communication. To remedy those marginal/borderline cases, following either FL or CPN/PB techniques, a score margin may be introduced marking a minimum difference on the candidate networks' scores and a hysteresis (i.e., time) margin to discourage very frequent VHO initiations [**113**]. The extent to which the score and hysteresis margins are changed to encourage or discourage a handoff depends on trends indicated through the values of relevant parameters. For example in [**114**], the authors used criteria such as the RSS-based link quality and the distance between

MN and base station and made use of training algorithms, proving that minimization of unnecessary handovers (approx. by 20%) can be achieved, optimizing the resulting performance by 10-20%, considering throughput, delay and packet loss.

3.2.5 Autonomic Features Addressing Robustness

Autonomic features could be exploited in various ways, towards enhancing the robustness of the VHO decision making. To begin with, awareness by definition aims at untangling uncertainties [**25**], resulting in enhanced robustness. Moreover, as already mentioned enhanced awareness considers also the frequency of parameters monitoring, which affects the precision of the parameter values that provide the basis for the uniform representation process. Along a similar line of reasoning, adaptivity and flexibility are essential for allowing the dynamic modification of the membership functions of FL systems, or the upper and lower values used by CPN approaches for the uniform representation of parameters, as well as, the score and hysteresis margins used to avoid marginal/borderline cases.

The aforementioned features can be beneficially combined with learning mechanisms (e.g., those based on neural networks) enabling the exploitation of historically available data and making use of memory strength, to optimize the tuning of upper and lower values, membership functions, and/or score and hysteresis margins and to help in combating more effectively context uncertainties or incompleteness.

Moreover, proactive measurements enable the analysis of the current state and the anticipation of events and their effect on the system, which would assist in addressing marginal/borderline cases. For example, estimations for the next possible location of the user, would fine-tune the hysteresis margin, preventing unnecessary VHOs resulting from the ping-pong effect.

TABLE III. Robustness Issues in VHO Decision Making and the relation with Autonomic Features

Robustness Issues		Awareness	Adaptivity & Flexibility	Learning	Proactivity
Diversity of parameters (Context diversity)	**CPN:** Different upper and lower values or thresholds for each parameter for each application.	Enables multi-layer parameter selection and defines the precision of the parameter values.	Enable the dynamic modification of upper and lower values or thresholds / membership functions / score and hysteresis margins.	Optimizes the tuning of upper and lower values or thresholds / membership functions / score and hysteresis margins and deals with uncertainties and incompleteness.	Anticipates upcoming events confronting marginal/borderline cases.
	PB: Different policy for each parameter for each application.				
	FL: Different membership functions for each parameter for each application.				
Diversity of criteria/ rules:	**CPN:** Different importance for each individual parameter, AHP.				
	PB: Policies' conflict resolution.				
	FL: Parallel fuzzy rules.				
Context uncertainties and incompleteness:	**CPN/PB/FL:** Verify the measured data through comparison with related data from other layer(s) to combat uncertainties. Substitute an average (**CPN/PB)** or neutral (**FL)** value (or, a value derived from historically available data) for the missing criteria.				
Marginal/borderline cases:	**CPN/PB/FL:** adjust score and hysteresis margins.				

4. A Comparison and Discussion on Selected Context-Aware VHO Management Solutions with Autonomic Orientation

4.1 Selected Autonomic-Oriented VHO Management Solutions

To demonstrate the applicability of the general concepts previously discussed, we now review six representative VHO management solutions with an autonomic orientation, taken from the literature. All reviewed proposals possess some context-awareness and cognitive characteristics, but differ in terms of the management architecture, the scope of information collection, the computational methods employed and/or possibly other aspects.

In the rest of Chapter 4.1 we individually examine each solution in turn, identifying relevant characteristics and associating them with the presented classification and taxonomy of Chapter 2; a summary of the results appears in TABLE IV. Subsequently, in Chapter 4.2 we compare the six VHO management solutions with respect to a number of criteria, presented in Chapter 3.

4.1.1 A Simple Terminal-Controlled Autonomic VHO Management Approach (TCAM)

TCAM [**95**] enables a simple and light-weight VHO management approach that does not require changes in the network infrastructure. All the intelligence lies in the MN and the handover is mobile-controlled, thus the type of the management architecture (see Chapter 2.2.4.2) is flat. Information collection (Chapter 2.2.1) is addressed by the MN, which monitors the RSS and SINR over the available radio interfaces, remaining energy level on the device's battery, and the velocity of the MN's motion. Velocity information is directly

deduced from the Doppler spread in the received signal envelope. User preferences are also included, considering QoS, monetary cost and energy efficiency, where the user asserts priority for each one.

The knowledge base (Chapter 2.2.2) includes a user profiles repository that contains the identities with which the user accesses different radio networks and the respective subscriptions to services, the user preferences and mobility policies. Additionally, the MN maintains a mobility policy database that contains a *black list of access network operators* with whom the user has had a bad experience. This feature enhances memory strength and enables learning. Both the user profiles repository, and the mobility policy database reside in the MN.

The handover decision process (Chapter 2.2.3) employs a set of parameters including both static and dynamic contextual information, and thus presents context-time variability. However, the parameters' set does not possess a high level of abstraction, as only physical-layer QoS parameters (SINR, RSS) are used. The parameters processing is based on a decision function, whose weights are dynamically adjusted according to user preferences. The system relies on users' criteria scoring for conflict resolution. Regarding the output format**,** the scheme produces one selected access network, having the highest score according to the user preferences.

4.1.2 An Autonomic VHO Scheme with a Client/Server Application Module (CSAP)

CSAP [**111**] has been claimed to be one of the first solutions that can function under diverse real-world scenarios involving a multitude of network technologies, network providers and applications. To achieve this versatility, the solution adopts a client/server scheme operating at OSI Layer 7 through a pair of applications: the *CNAPT (Client Network Address and Port Translator)*, which resides at the MN, and the *SNAPT (Server Network Address and Port*

Translator) at the network side. These applications abstract technology-dependent details and introduce a form of virtualization.

The management architecture is described as having an adjustable structure, a characterization stemming from the versatile form of cooperation between the CNAPT at the MN and the SNAPT at the network side. In view of this fact, the VHO management architecture of this solution can be classified as hybrid.

Considering information collection**,** user, terminal and network context is gathered by the CNAPT, with assistance from the SNAPT. It is stated that the system periodically searches for available network connections (search activity) and at the same time, periodically verifies reliability and performance of the current connection (check activity). The check activity is related to sampling of the RSS at the physical layer and to application-layer parameters, inferring the experienced Round-Trip-Time (RTT) with the help of ping messages. The scheme does not consider monitoring of variables at the link-layer, since some NICs do not support reading such values through standard APIs. With respect to the knowledge base, the CNAPT includes a history repository, providing a high memory strength.

The parameters selection step of the handover decision making presents context-time variability, as it includes not only static, but also dynamic contextual information (RSS, RTT). A higher level of abstraction is supported in comparison to TCAM, as both physical and application layer QoS parameters are considered. These give some indication on the effective status of the connection (i.e., being active or not) and of the effective load. Still, the considered parameters do not span all layers. Concerning parameters processing, handover initiation and network selection processes are based on a generic framework based on thresholds, which can be classified as a form of PB processing. Specifically, if the reliability or performance index goes below the specified critical thresholds or the current network connection is experiencing an interruption, the ‘check’ activity triggers the handover

initiation procedure. The ensuing network selection relies on the results provided by the search activity.

4.1.3 An Intelligent Cross-Layer Terminal-Controlled VHO Management Scheme (CLTC)

CLTC [**107**] is another mobile-controlled handover scheme with a flat VHO management architecture, placing all intelligence on the mobile devices. The information collection is implemented by the MN through monitoring and measurements, to identify the need for handover. The context information can be relative to the network, the terminal, the service and the user. QoS parameters are included, such as bandwidth, delay, jitter, packet loss, RSS and BER of the current access network and the neighboring available networks. Furthermore, context information related to user preferences, service capabilities (real-time and non real-time), MN status (battery and network interfaces), priority given to interfaces, location and velocity is collected. The knowledge base includes a policy repository maintained in the MN, but this repository does not provide support for the assessment of past policies and VHO decisions.

The parameter selection step of the handover decision making is dynamically adjustable, determined by multiple criteria. The selected parameters present context time variability, including both static and dynamic contextual information (such as access network availability, MN's velocity, etc.). Moreover, a high level of abstraction is supported, as QoS-related parameters are extracted from all network layers. With respect to the parameter processing, VHO initiation employs FL. The information gathered is fed into a fuzzifier converting the aforementioned elements into fuzzy sets. A fuzzy set contains a varying degree of membership in a set. For instance, RSS can be weak, medium or strong. After fuzzification, fuzzy sets are fed to an inference engine, where a set of fuzzy rules are applied to determine whether the handover is necessary. Fuzzy rules utilize a series of IF-THEN rules

and the result is YES, Probably YES, Probably NO or NO. At the final step, the resultant decision sets have to be "defuzzified". For that, the centroid method is used to obtain a handover initiation factor (YES or NO) based on membership values and decision sets.

If a handover is necessary, the network selection stage is based on an AHP method that allows the decomposition of the network selection problem into several sub-problems, corresponding to the decision criteria. The method assigns a weight to each sub-problem and calculates for each network the weighted sum characterizing the cumulative impact of all criteria. The output format of the parameter processing process is a ranked list of handover targets, with networks featuring higher weighted sums placed closer to the top of the list.

4.1.4 PROTON: An Autonomic VHO Framework with Finite State Transducers

A primary characteristic of PROTON [**66**] is a metric called TF, related to a Finite State Transducer with Tautness Functions and Identities (TFFST), which enables policy modeling and resolves potential conflicts. The relevant management architecture includes components at both of the network and MN sides. Those components that involve heavy computations are placed on the network side, to minimize complexities at the MN. The VHO is initiated by the MN, but uses assistance from the network side, which provides information and computational services (the TFFST models creation). Thus, the overall handover process can be classified as network-assisted. Since the main managing entity controlling the handover process is on the MN, the management architecture is distributed hierarchical.

The information collection activity is implemented by the ‘sentinels’ and ‘retrievers’, located at the terminal. The sentinels are responsible for collecting dynamic elements, whereas the retrievers manage static elements (e.g., user preferences or application profiles). The knowledge base includes a policy repository on the network side and a TFFST Repository on the MN. During the handover decision making, the parameter selection process is driven by policies and it is divided into three steps, executed on the MN. Specifically, the

collected information is filtered according to simple local rules, and then it is grouped into sets. The parameters used include both static and dynamic context originating from the physical, network and application layers.

The parameter processing occurs on the network side, where the conflict resolution module builds a deterministic Finite State Machine modeling every active policy, and subsequently generates the set of TFFST profiles, which is flexible and can be updated according to the MN requirements. During the TFFST profiles generation, all possible static and dynamic conflicts are foreseen. Therefore, the algorithms that are executed have a high computational cost. Subsequently, the mobile device stores and uses the TFFST profiles, to be able to react quickly to incoming events. In order to prioritize TFFST profiles, the tautness function is formed, to indicate how tautly a condition fits to an event. In order to quantitatively represent the tautness, a real number in the interval [−1, 1] is used, so that the stronger a condition is, the closer its TF is to zero. The corresponding output format is the most fitting candidate network.

4.1.5 An Autonomic VHO Approach with a Context Evaluation Matrix at the Network Side (COEVAL)

COEVAL [**73**] implements VHOs by introducing a context evaluation matrix and a respective context evaluation function. The management architecture may be classified as distributed hierarchical, as it includes cooperation between terminal and network side components. More precisely, the context server located in the network collects information, compiling it into a matrix, in response to the handover initiation request from the MN. During the subsequent network selection the MN processes the matrix and makes the handover decision, also considering current dynamic information. In view of these facts, the handover control method can be characterized as network-assisted.

Considering information collection, the scheme provides the mechanisms for the collection, aggregation and filtering of contextual information, utilizing context from the MN and the context server (located in the network. More precisely, the context server collects the relevant context information from the various context repositories. Then, the MN collects dynamic context such as the received signal strength, the CPU usage and the remaining charge on the battery and combines the collected information with the data from the context server.

With respect to the knowledge base, the main entity is the context server (that has no memory of past events), in addition to various other context repositories, including the respective Operations and Support Systems (OSS), the location information database and the user profile database, all of them residing at the network side.

During the handover decision making, the parameter selection is based on the received information from the context server and the current dynamic information from the MN, derived from all the layers. The parameter processing method uses a context evaluation (decision) function that manipulates the matrix context using dynamically adjustable weights and chooses the appropriate network interface for each application, taking into account both user and network preferences. The output of the context evaluation function is the appropriate network interface for each running application.

4.1.6 AUHO: An Autonomic Personalized Handover Decision Scheme

AUHO [**65**], employs the same architecture, information collection process and knowledge base components with COEVAL, in conjunction with a different decision making process, which employs FL and MAD. Specifically, the parameter selection step of the handover decision making employs contextual information provided from all the layers, including dynamic context.

TABLE IV. Classification of the selected VHO management solutions, according to their characteristics

Solution	Terminal Side Functionalities	Network Side Functionalities	Type of Management Architecture/ Handover Control	Parameters Processing Method
TCAM [**95**]	All the intelligence at the MN, VHO Decision & Execution	None	Flat / Mobile Controlled	DF, with dynamically adjustable weights (based on user preferences)
CSAP [**111**]	VHO Initiation, Network Selection & Execution	Assists the MN during the information collection and the network selection (search & check activities)	Hybrid / Network Assisted	Generic PB framework based on thresholds
CLTC [**107**]	All the intelligence at the MN, VHO Decision & Execution	None	Flat / Mobile Controlled	FL for VHO initiation & AHP (i.e., MAD) for network selection
PROTON [**66**]	VHO Initiation & Network Selection (TF computation), VHO Execution	Computationally demanding tasks, TFFST profiles computation	Distributed Hierarchical / Network Assisted	Policy-based: TFFST model creation, implemented with Finite State Automata & Tautness Function
COEVAL [**73**]	VHO Initiation, Network Selection & Execution	Context Server and various context repositories assist in information collection	Distributed Hierarchical / Network Assisted	Context Evaluation Function (i.e. DF)
AUHO [**65**]	VHO Initiation, Network Selection & Execution	Context Server and various context repositories assist in information collection	Distributed Hierarchical / Network Assisted	FL for VHO initiation & Additive Aggregate Utility Function (i.e., MAD) for network selection

Additionally, user preferences are perceived, as a set of attributes ordered from most to least desired, considering RSS, Cost, Quality and Lifetime (i.e., remaining battery charge). Considering parameter processing, the handover initiation stage is performed by means of a FL-based method, employing fuzzification and defuzzification mechanisms for the

calculation of APAV (Access Point Acceptance Value) for all available networks. The network selection employs an additive aggregate utility function (i.e., a MAD function), which computes the APSV (Access Point Satisfaction Value) for all candidate networks and chooses the most satisfying network. The output format is formed by choosing among the best access points (based on RSS, Quality, Cost and Lifetime) the one being most important to the user (prioritized set of candidates), for each application.

4.2 Comparison and Discussion

We now compare the presented VHO solutions according to the extent these solutions incorporate and exploit the autonomic features of Chapter 3.1, towards enhancing the effectiveness and efficiency of VHOs. The comparison also addresses the robustness issues identified in Chapter 3.2.

Additionally, we consider issues related with the operational complexity. This is another important aspect, which determines the achievable degree of self-management for the MN. Operational complexity can be generically characterized as the "degree of complexity of memory and time" [**28**] and can be linked to the computational overhead and the signaling overhead. Accordingly, the comparison of the solutions also considers the tradeoff between intelligence/ sophistication and operational complexity. A summary appears in Table V.

4.2.1 Considering Autonomic Features

4.2.1.1 Awareness

Awareness, the basis of all other autonomic criteria, is related with the information collection and parameter selection processes. All aforementioned VHO solutions present awareness, though in a varying degree: the two simpler and more lightweight approaches, namely TCAM and CSAP, provide a basic form of awareness, while the other solutions exhibit more enhanced awareness, but at the cost of higher complexity. Specifically, TCAM

limits information collection to just physical layer parameters used to measure the signal quality of the candidate network (SINR, RSS). Thus, TCAM does not have potential for multi-QoS consideration. CSAP takes a simple approach too, but supplements the physical layer monitoring of RSS with the application layer monitoring of RTT (Round Trip Time), which gives some indication on the effective status of the connection, the effective load and the available throughput. Still, the level of abstraction is not high enough to provide potential for explicit multi-QoS consideration.

Turning to the more sophisticated approaches, CLTC implements active monitoring, with parameters extracted from all layers. The solution considers QoS parameters (bandwidth, delay, jitter, packet loss, traffic load), coverage, monetary cost, link quality (RSS and BER) of the current access network and its neighbors, as well as location information. However, all these parameters may not be needed in every scenario, thus, the tradeoff between enhanced awareness and the resulting signaling and computational overheads should be taken into account, especially considering that according to this approach all the intelligence is placed at the MNs.

PROTON provides active monitoring and a high degree of awareness, through monitoring parameters at different layers and organizing the collected data according to a three-level hierarchy, which reduces the volume of data processing. While PROTON's framework could in principle enable multi-QoS consideration, it lacks information necessary for explicitly considering the demands of different running applications. This shortcoming might be due to the fact that PROTON is one of the first approaches on autonomic VHO management.

AUHO and COEVAL support a high degree of awareness through active monitoring. More precisely, the mobile device performs measurements to retrieve updated dynamic and static information from its sensors, the user and the context server (in the network side). Emphasis is given to information related to the running application requirements and the available

network interfaces, considering parameters such as bandwidth, packet error rate, delay, jitter and packet loss ratio, which assist in proposing the best network interface for each running application. Also, location information is included in the handover management parameters that add a spatial dimension in the handover initiation criteria.

4.2.1.2 Adaptivity and Flexiblity

Considering adaptivity and flexibility, the information collection mechanisms and handover decision making processes are compared in view of the presented solutions. In general, there is an inherent trade-off between flexibility and computational overhead. TCAM, being the simplest and most lightweight solution, is characterized by a rather limited adaptivity and flexibility, as it deals with a predefined set of parameters and does not provide adaptation mechanisms in information collection. The other approaches present more enhanced adaptivity and flexibility characteristics, but are also more computationally demanding. In CSAP, for example, the monitoring activity is still non-adaptive, as it uses a constant rate of parameters sampling. However, adaptive thresholds are used in parameter processing ("check activity").

In CLTC, the Analytic Hierarchy Process offers advanced flexibility and also adaptivity (through the possibility of dynamically adapting the various weighting factors). On the other hand, adaptive monitoring mechanisms are not considered. By contrast, PROTON offers sophisticated monitoring adaptivity, as each parameter is collected according to a specific polling frequency, depending on connectivity resources and mobility profiles. Specifically, the system adapts the frequency of active monitoring proportionally to the MN's velocity, matching thus the information collection rate to variations of the user's physical mobility. Considering the decision phase, PROTON, provides advanced flexibility and accordingly provides enhanced adaptivity mechanisms through a dynamic set of TFFSTs and the use of suitable tautness functions for each case.

In COEVAL, the MN fills in the dynamic contextual information and calculates the evaluation matrix when a decision is needed, applying policies that may include rules to set the upper or lower bounds. The matrix mechanism provides advanced flexibility, being dynamically filled with the available parameters. Also, the dynamic upper/lower bounds offer adaptivity. Finally, AUHO features advanced flexibility through the Multiple Attribute Decision method, where the output is calculated as a linear function of context input and dynamically changing weights, with respect to different criteria.

4.2.1.3 Learning

This autonomic criterion is related to the ability of the system to learn, enabled by the memory strength provided by historically available data. Interestingly, only the simpler approaches, TCAM and CSAP, provide a form of memory strength, which can be exploited to include learning mechanisms. In TCAM a black list of access network operators is included, containing the networks where the user has had a bad experience.

Additionally, the description of the solution [**95**] mentions that users can specify and alter their preferences dynamically, through a learning process. CSAP employs a repository of the most significant information, containing trends, failures, trajectories, user choices, etc., about past experience, and providing the ability to adjust internal parameters and derive statistical measures of trend.

For instance, if an on-board GPS is available, the system can decide to store and learn maps identifying good coverage areas together with the characteristics of the network access that can provide the coverage. This might prove quite useful in the case of users constantly traveling along the same routes, as it is the case for people daily commuting between their homes and work places.

4.2.1.4 Proactivity

Proactivity is based on preventive measurements promoting actions in the direction of system anticipation. The solutions under investigation that involve proactive mechanisms are CSAP, CLTC and PROTON. Specifically, the description of CSAP [**111**] mentions that the system is able to efficiently smooth the sampled values of measures RSS, through simple weighted moving averages, and at the same time calculate a simple trend indicator to be used in cross-validation with the moving average, enhancing proactivity.

Considering CLTC, in a new and enhanced version of the approach [**72**], predictive Link layer information is taken into account extending the proactivity of the solution. More specifically, the system detects the quality of the current link (concerning physical and MAC layers) and can issue periodically a polling command to check the status of the link, expressing the likelihood of future changes in the link properties (e.g., link going down, link going up, etc) based on present conditions. Finally, PROTON implements a conflict resolution module to resolve conflicts among the policy rules. During this task, all possible static and dynamic conflicts are foreseen, enhancing proactivity.

TABLE V. Comparison criteria for autonomic VHO management schemes

Autonomic Criteria	Awareness	Adaptivity & Flexibility	Learning	Proactivity	Robustness
TCAM	low level of abstraction: only physical layer parameters to measure the signal quality	limited flexibility, non-adaptive monitoring, limited adaptivity in the decision making (dynamic weights applied for user preferences only)	black list of access network operators	no information about the frequency of monitoring measurements, relies only on users' criteria scoring for conflict resolution	no multi-QoS consideration, no provision for managing uncertainties or incompleteness, no marginal/borderline cases consideration

CSAP	medium level of abstraction: physical (RSS) & application (RTT) parameter selection	advanced flexibility, adaptive thresholds in decision making, non-adaptive monitoring	repository of the most significant information (trends, failures, trajectories, user choices) about past experience	"check activity" follows a periodic activation scheme, trend indicator	no full specifications provided about the network selection process; temporary fluctuations of physical parameters are verified by application layer ping messages, managing uncertainties and avoiding ping-pong effect
CLTC	high level of abstraction: multi-layer parameter selection	advanced flexibility, adaptive decision making (through adaptive weights), non-adaptive monitoring	no memory strength	predictive link layer information (in the new version [**72**])	enhanced robustness: context and criteria diversity consideration through parallel fuzzy rules and AHP, multi-QoS consideration, provision for managing incompleteness, but no marginal/borderline cases consideration
PROTON	high level of abstraction: multi-layer parameter selection, three-level organizational hierarchy of collected data	advanced flexibility, adaptive decision making, well structured adaptive monitoring framework	no memory strength	all possible static and dynamic conflicts are foreseen	no multi-QoS consideration, criteria diversity consideration through policies' conflict resolution, policies related to hysteresis that could manage marginal/borderline cases
COEVAL	high level of abstraction: multi-layer parameter selection	advanced flexibility, adaptive decision making (adaptive thresholds), non-adaptive monitoring	no memory strength	no	multi-QoS consideration, no provision for managing uncertainties/incompleteness, no marginal/borderline cases consideration
AUHO	high level of abstraction: multi-layer parameter selection	advanced flexibility, adaptive decision making, non-adaptive monitoring	no memory strength	no	enhanced robustness: context and criteria diversity consideration through parallel fuzzy rules, multi-QoS consideration, provision for managing incompleteness, but no marginal/borderline cases consideration

4.2.2 Considering Robustness Issues

We now focus on the robustness issues considering the decision making procedure. With respect to the uniform representation enabling context diversity consideration, TCAM and COEVAL use CPN, while, CLTC and AUHO use a combination of FL and CPN techniques. Lastly, CSAP and PROTON use PB techniques, where each parameter is processed individually, through a relevant policy.

Adjustable tailoring of the normalization parameters, according to QoS demands for each running application is considered in COEVAL, but not in TCAM. CLTC and AUHO follow the framework proposed by [**115**] and consider different membership functions for each application, while they use CPN to simpler criteria. CSAP and PROTON lack information about multi-QoS considerations according to running applications requirements.

Considering the diversity of criteria/rules, including the assignment of a different importance to each group of criteria, the most comprehensive approach is taken by CLTC and AUHO, which use FL with parallel fuzzy rules. Additionally, in CLTC parameters are grouped into a hierarchical model, in order to be handled more efficiently through AHP.

However, this approach uses a rather complex weighting method, so the complexity of CLTC is higher than that of AUHO. COEVAL deals with the matter in simpler terms: while it allows the assignment of a different importance to each individual parameter, it does not provide support for handling an entire group of criteria. PROTON employs its policies conflict resolution module to combat diversity of criteria/rules. Finally, TCAM and CSAP inherently lack capabilities for dealing with complex decisions.

We now turn to the management of uncertainties and incomplete information during the decision making process. To combat incompleteness, the FL-based AUHO and CLTC solutions substitute a neutral value in place of missing parameters. PROTON, COEVAL and TCAM do not provide any explicit support for managing incompleteness.

Finally, CSAP provides some means to guard against uncertainties arising from excessive parameter value fluctuations. Specifically, the solution tries to avoid improper reaction to temporary fluctuations of physical layer parameters, by cross-checking a bad link status against application layer information (obtained through ping messages).

Concerning marginal/borderline cases and the ping-pong effect, predictive link layer information is included in the new version of CLTC [**72**], which could possibly assist in the confrontation of this problem, while, PROTON presents policies related to hysteresis margin. CSAP may deal with the ping-pong effect through its previously mentioned mechanism for managing uncertainties. The rest of the solutions do not include support for managing marginal/borderline cases. As a whole, the solutions with the most comprehensive provision for robustness are CLTC and AUHO.

4.2.3 General Comments towards Self-Management and Autonomicity

While the simpler solutions TCAM and CSAP provide only moderate potential for overall self-optimization, due to their incomplete awareness and limited adaptivity and flexibility, the overall complexity of the corresponding VHO decision making procedures is low, signifying a high degree of achievable self-management for the MN.

On the contrary, the performance potential of CLTC, PROTON, COEVAL and AUHO is greater, in view of their enhanced awareness and adaptivity & flexibility features, but this comes at the cost of a higher complexity. The complexity of CLTC, in particular, may be characterized as quite high, so the flat and mobile controlled architecture of this solution might prove impractical, as the heavy programming tasks could overwhelm the MNs.

PROTON, COEVAL and AUHO are better positioned in this respect, as their distributed hierarchical architecture and network assisted control foresee centralized entities to take up the heavy programming tasks, reducing the burden put on the MNs.

Another noteworthy aspect, particularly in a FI context, relates to the concurrent exploitation of different network interfaces on the MN for serving different running applications. In this direction, the parameters selection set should allow information from multiple layers to be included and matched with the running application requirements, so that the system can select the most appropriate access network for each running application. CLTC, AUHO and COEVAL provide the most comprehensive support for this.

As already mentioned, only TCAM and CSAP provide some form of memory strength that may be exploited towards cognition and learning. This existence of memory strength might be seen as a supplement to the moderate degrees of awareness, flexibility and adaptivity present in these simpler solutions.

However, the other four more sophisticated solutions could also benefit from memory strength and additional learning mechanisms. Although the incorporation of such mechanisms may involve initially increased computational overheads, the more effective prevention of unnecessary VHOs could counter-balance these overheads and eventually lead to enhanced performance.

The additional mechanisms could be hosted by higher level entities, particularly for hybrid or hierarchical architectures, such as those in AUHO and COEVAL, avoiding an extra burden on the MNs. In AUHO, for example, pre-calculated APSV values characterizing the network interfaces under typical patterns of context could be stored in the knowledge base, towards faster and less computationally demanding decisions.

Along further directions, learning mechanisms can be employed to optimize the formulation of membership functions for FL-based solutions (AUHO and CLTC), to tune the upper/lower values and thresholds used for CPN (in COEVAL and TCAM), to optimize the formulation of policies in PB systems (CSAP and PROTON), or to formulate and optimize the score and hysteresis margins used when dealing with marginal/borderline cases.

5. Modeling Methodology for Performance Evaluation

Proper assessment of the relative merits of alternative VHO approaches requires a sufficiently comprehensive and generally applicable performance evaluation methodology. However, the methodologies available in the literature (reviewed in Chapter 5.2) are still limited, being either too simplistic to accurately capture the process, dynamics and sources of the context-related information dissemination, or too detailed and scenario-specific to be applicable for a comparative performance evaluation across architectural alternatives.

Towards addressing this gap, the dissertation provides a versatile modeling methodology that focuses on signaling in the VHO preparation phase and incorporates all significant aspects that are associated with the exchange, queueing and processing of the signaling messages and have an impact on delay-related performance, as presented in the rest of this Chapter. The resulting model is comprehensive, yet capable of producing closed form results. More importantly, the modeling methodology is generic and can be flexibly tailored to the characteristics of different VHO architectures, properly accounting for differences in the process of obtaining context-related information in each case.

This versatility is demonstrated through an application of the modeling methodology in two VHO architectural approaches that differ in their way of collecting dynamic resource availability context, presented in Chapters 6 and 7. In line with the previous discussion, the fist case follows standards-based recommendations and includes a CS that can handle only static context. Dynamic resource availability context is obtained reactively, during the processing of each handover. In the following, this case will be referred to as “On-demand Resource Information Gathering” (ORIG). The other considered case reflects recent proposals for architectural amendments and includes a local CS, coined "Dynamic Context

Repository" (DCR), that gathers proactively and periodically resource availability information (which varies, depending on the current traffic load) from a number of RANs associated with the DCR. This approach is called “Proactive Resource Information Gathering” (PRIG). For both cases, the analytical results are validated against simulations, presented in Chapter 8, addressing an appropriately rich set of relevant parameters, towards confirming the effectiveness and the accuracy of the modeling methodology.

5.1 Basic characteristics of major VHO management frameworks and related amendments

As already mentioned in the introduction in Chapter 1, major standardization bodies have provided specifications relating to VHO management frameworks. Most notable are the IEEE 802.21 Media Independent Handover Services [**12**] – recently updated with the IEEE 802.21.2017 [**13**] and IEEE 802.21.1.2017 [**14**] Media Independent Services Framework standards and the 3rd Generation Partnership Project (3GPP), which proposed the Access Network Discovery and Selection Function (ANDSF) Management Object (MO) [**15**].

These specifications describe the mechanisms and operator-defined policies, by which an entity may discover and obtain contextual information about a (possibly heterogeneous) set of networks serving the entity's geographical area, for use in network selection decisions. Both specifications address all types of handover control (i.e., mobile controlled, mobile assisted, network assisted, and network controlled handovers) and include a CS that stores all the static contextual information and the associated policies. This CS is called the "Media Independent Information Server" (MIIS) in IEEE 802.11 and the "ANDSF Server" in the ANDSF specification. The information provided by the CS includes a list of the available access networks and discovery information and policies according to operator requirements. The CS

may also include static link layer parameters, such as channel information, roaming agreements between different operators, costs for using the network, etc.

From a more specialized viewpoint, but still relevant to a seamless VHO management, the Hotspot 2.0 standard from Wi-Fi Alliance [**116**] improves the ability of WLAN stations to discover and connect in a secure way to public Wi-Fi access points (APs). Hotspot 2.0 builds on top of the IEEE 802.11u specifications [**117**] that enable devices to discover information about the available roaming partners, using query mechanisms capable of collecting contextual information. Additionally, 3GPP [**118**] provides alignment and complementarity of ANDSF and HotSpot 2.0 policies, which can be leveraged towards supporting a number of multi-operator scenarios [**16**], [**119**], [**120**].

As noted in the introduction, the aforementioned specifications provide support only for static context. Important dynamic context (e.g., traffic load levels, or the availability of resources in a given CN, also dependent on traffic load variations) is collected reactively each time a handover is triggered (see, e.g., [**34**]), through the direct interaction between the MN (or the hosting SN) and each CN involved. The heavy volume of signaling required for this process can be avoided if the required dynamic context is proactively gathered, in a periodical fashion. To enable such an approach, the standards-based VHO frameworks must be amended or extended, to provide support for obtaining and storing the dynamic context. Such amendments have been proposed by [**17**], [**18**], [**19**], [**16**].

Specifically, [**18**] introduces an enhanced CS that receives regularly dynamic contextual information (e.g., the available bandwidth) from all RANs in its domain. The main limitation of this proposal is that the enhanced CS remains a single centralized component that would have to monitor the status of a potentially vast number of RANs. This arrangement could

pose significant complexity and scalability issues. For this reason, the architecture proposed in [**17**], apart from the centralized (static) CS includes also a dynamic CS per radio access technology, which gathers contextual information updates from the RANs employing this technology. The work [**16**] proceeds along similar lines, this time leveraging the integration between ANDSF and Hotspot 2.0. Specifically, the UEs collect information from a local instance of ANDSF ("Local ANDSF", as in, e.g., [**119**], [**120**]) about the policies of the operator for accessing the various RANs in the area, as well as dynamic information from Hotspot 2.0 protocols, to evaluate the status of WiFi APs (e.g., number of users associated to the AP, the load on the backhaul link etc.). An even more disruptive model (in comparison to the standards-based frameworks) is presented in [**19**], eliminating altogether the centralized static CS defined by the standards and replacing it with an architecture possessing three layers of hierarchy. The hierarchical structure involves hash tree-based information servers, which, instead of storing the full data, they register a reference that points to the corresponding load-aware server where the respective data is stored.

The PRIG case studied in this dissertation incorporates aspects from the amendments just reviewed, by retaining the single centralized CS for the static context and introducing local CSs (the DCRs) to handle dynamic context. Each DCR is responsible for a number of (possibly heterogeneous) RANs in its local area. By making this number larger, increasingly more centralized configurations (closer in spirit to the approach in [**18**]) are obtained, with fewer DCRs managing more RANs.

5.2 Methodologies for the performance evaluation of VHO frameworks

Considering the aforementioned access network selection frameworks, it is important to assess each proposed VHO management model, in order to provide insight about the performance indicators, such as end to end latency, signaling overhead etc. Quantitative assessments that exist in the literature can be separated in three categories.

The first, includes approaches with crude estimation of end-to-end latency, aggregating successive mean time intervals for message exchanges and identifying which steps in the message sequence may cause more delay overhead than the others, without taking into account queueing phenomena. Examples of such approaches can be found in [**17**], [**19**], [**121**].

The second category includes brute force simulations of the proposed architectures [**17**], [**19**], [**16**], taking into account the rate of handover triggers, considering in a more realistic way the wired and wireless link delays. However, such approaches of simulation evaluations are valid only for specific heterogeneous network scenarios and it is difficult to extrapolate their outcomes in order to compare different approaches.

The third category includes analytical model based evaluations, with the objective to capture queueing phenomena that are caused by the assumed network topology and the related signaling. Model based evaluations are capable to be modified to adapt to framework changes, and thus can be used by different frameworks and their variations. In [**18**] the authors provided an analytical framework, having included as parameters the intensity of handover triggers, the number of MNs, link parameters and queueing phenomena assumed within the topology of the basic architectural components.

Yet the evaluation model did not quantify other additional important factors, such as the possibility of requiring more than one networks to be checked until finding a suitable network target, which was incorporated in our previous work [**34**]. More specifically, in [**34**] we provided an analytical model that focused on a reactive resource information gathering scheme, concluding that signaling overhead and end-to-end delay are heavily affected from the intensity of the VHO requests, the number of MN users in a RAN, as well as, the resources availability probability of a RAN. The latter is due to the fact that when it is harder to find available resources, more networks have to be queried sequentially, and thus more traffic is generated throughout the network segments, resulting in additional load in the queues.

In the following, we provide an analytical system model, extending the work presented in [**34**], redefining the analytical methodology to be generic and flexible enough, in order to be easily applied in various cases of network architectures. The system model is comprehensive, yet able to produce closed form results.

More specifically, the proposed modeling methodology is used to compare diverse architectural approaches that present different strategies of checking the resource-related information (i.e. reactive or proactive), demonstrating also the impact of computational resources scaling on the overall end-to-end delay, in order to prove the feasibility of each approach considering efficient RAN selection in next generation networks. Comparison with simulation confirms the accuracy of analytical results.

5.3 Elements of the Modeling Methodology

5.3.1 Topological and Architectural Considerations

The generic system model addresses an environment involving N_{RANs} distinct RANs serving a given area. These RANs may employ the same or different radio access technologies. Each RAN acts as the SN for a number of MNs. When a handover is triggered for an MN, the set of other RANs in the area that qualify as CNs for this MN must be identified. Subsequently, when a particular CN in this set is considered as the handover target, the availability of the necessary amount of resources therein must also be checked. To assist with these tasks, the VHO architecture includes at least one CS, which provides a repository of static data and policies to determine the set of CNs. Addition-al CS(s) (or extensions to the static CS) may also be available, with facilities to collect and store dynamic context for use in the resources' availability check. If such advanced capabilities are not present, the availability of resources is checked through direct queries to each of the CNs examined.

The operations just outlined involve the exchange of signaling messages between components of the VHO management architecture. Along this process, the SN of the MN subject to handover acts as a mediator between the MN and other entities (CNs or CS(s)). Signaling between different RANs or between a RAN and the CS(s) occurs over wired links of the backhaul, while signaling between a MN and its SN occurs over the RAN's wireless link. Figure 8 provides an outline of the characteristics just discussed.

In the interest of presenting the essential elements of the modeling methodology as straightforwardly as possible, in the following we assume a homogeneous setup, in which all

RANs serve an equal number of MNs each, denoted as N_{MN}. Moreover, it is assumed that the set of CNs for an MN subject to handover always includes N_{CN} RANs and that any of the $N_{RANs}-1$ local RANs besides the SN are equally likely to belong to this set. Finally, all RANs are taken to have the same wireless link characteristics and all links in the wired backbone are assumed to have the same capacity. All these simplifications can be relaxed and the results can be readily adjusted for addressing a heterogeneous setup, at the expense of somewhat more complicated formulas for the results and some extra notation to express the asymmetries.

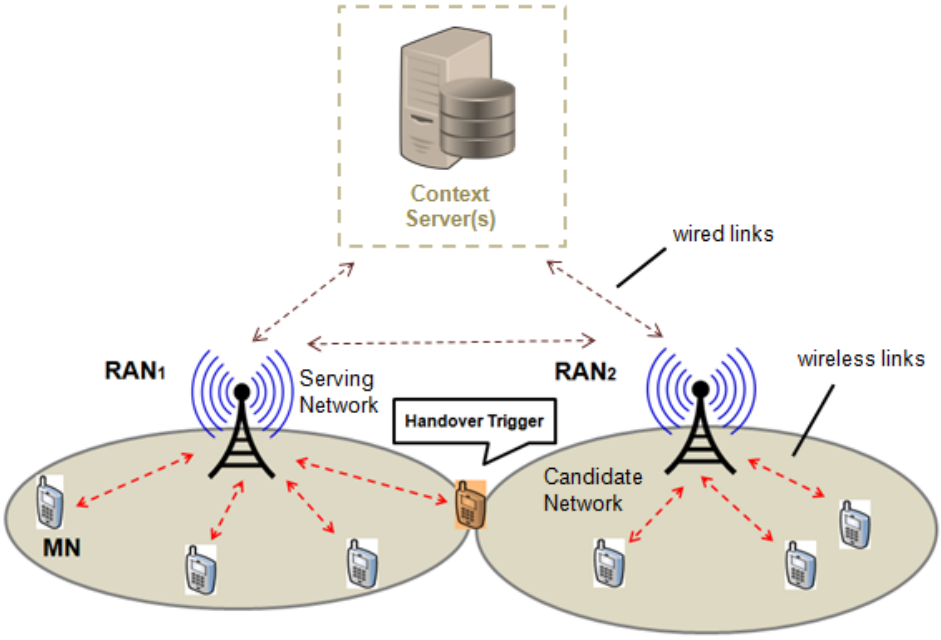

Figure 8. Main entities concerning the generic system model.

5.3.2 Interactions between components of the VHO architecture and related signaling

We now provide a more detailed description of the individual steps in the VHO preparation process, outlined in the generic message sequence chart (MSC) of Figure 9. The steps in this figure are in accordance with the Mobile-Initiated HandOver (MIHO) scenario, in which the handover trigger originates from the MN. With simple adjustments, the MSC can also accommodate the complementary Network Initiated HandOver (NIHO) scenario.

The VHO preparation process, which is one of the most critical phases to control during the whole VHO process, due to its complexity [**17**], includes two phases, corresponding to the two shaded areas in Figure 9. The VHO preparation process involves the related signaling initiated with the handover trigger and completed with the selection of a suitable network target. The first phase (shaded area 1) begins with the VHO trigger and ends when the MN (or its SN) receives a list of CNs. In more detail, when the conditions for a handover trigger are met, the MN issues a VHO request to its SN (message a in Figure 9).

Upon reception of this message, the SN acts as a mediator and retrieves information specifying which RANs in the area are suitable as CNs. The relevant contextual information is provided through a CS hosting static context and policies, as foreseen in major standards (IEEE 802.21 and its evolutions, or 3GPP ANDSF). To implement the aforementioned activity, the SN forwards the VHO request (message b) to the CS and receives in reply (message c) a sorted list of RANs that qualify as CNs for this MN.

The contents of the list are determined on the basis of available static information, such as the supported data rate(s) of each radio access technology, coverage, pricing, nominal energy consumption, etc. In principle, the first phase is complete once message c with the list of CNs is received at the SN. However, depending on the specification of the VHO preparation

process, this phase may also involve two additional messages: one message to forward the list of CNs from the SN to the MN (message d), and a subsequent message from the MN to the SN (message e) to initiate the resource availability checks. The potential inclusion of these two final messages is shown at the bottom of shaded area 1 in Figure 9.

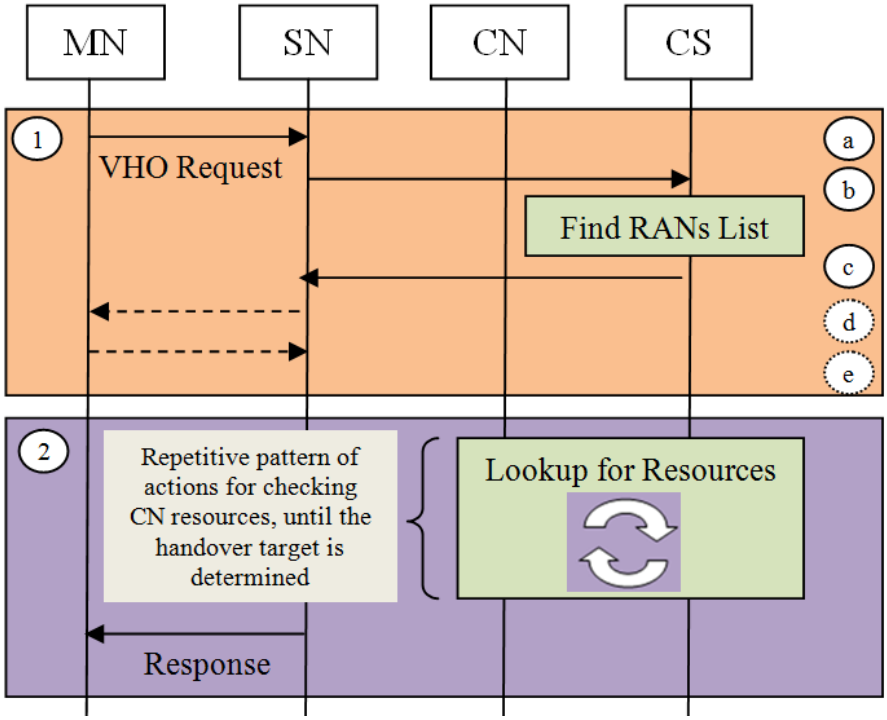

Figure 9. A generic MSC depicting the VHO preparation procedure.

The second phase includes the actions required for determining the handover target among the CNs in the list obtained from the first phase. For this, the CNs are examined one by one, in the order listed, checking if the residual amount of available resources at the CN suffices for admitting the MN subject to handover. Once a CN with adequate resources is encountered, it becomes the handover target and the handover preparation process ends with a VHO response message sent from the SN to the MN (depicted at the bottom of shared area 2 in Figure 9).

In case all CNs in the list have less than adequate resources, the handover fails and the message sent to the MN includes a negative response. In view of these remarks, the actions required for checking the availability of resources in a CN occur repetitively, until the network target is determined (or the list of CNs is exhausted), and the iterative nature of the process is shown in Figure 9. The number of required iterations is a random variable, whose distribution depends on the number of CNs and on the likelihood that an examined CN will be found having adequate resources. Further properties of this distribution are discussed in the next subsection.

In view of its generic character, the MSC in Figure 2 omits the signaling required for checking the availability of resources in the CNs, because the details of this signaling depend on further properties of the VHO management architecture. Thus, the generic MSC must be expanded to comply with the particular architecture under study (as done in Chapters 6 and 7).

Specifically, in architectures without support for dynamic context, the SN must examine separately each CN, by sending to it a query message and receiving the corresponding reply. Moreover, this ex-change of messages must be replicated iteratively, as indicated in Figure 9, to reflect the sequence of checks per-formed until the handover target is determined. Architectures with dynamic context support have more modest signaling requirements: The SN sends a single query for all CNs to the CS managing the dynamic context, which performs the required checks and replies indicating the handover target. In this case, the iterations shown in Figure 9 do not refer to additional signaling, but to the length of time required for processing the query at the CS.

5.3.3 Distribution of the number of CN checks

As already mentioned, the required number of checks R is a random variable, whose distribution depends on the number of CNs N_{CN} and on the likelihood that an examined CN will be found having adequate resources. This likelihood is expressed through the probability p, which quantifies the congestion in the CN being checked (RANs with higher load corresponding to lower values of p). Aligning with the homogeneous nature of the setup considered, the same value of is employed for all CNs.

In view of these characteristics, the total number of CN checks R is the number of steps in a sequence of independent Bernoulli trials, until encountering the first success (occurring with probability p) or until completing N_{CN} steps (this event corresponding to checking all CNs in the list without success). Thus, the distribution of R has the truncated geometric form

$$\Pr\{R=n\}=\begin{cases}(1-p)^{n-1}p, & 1\le n<N_{\mathrm{CN}},\\ (1-p)^{N_{\mathrm{CN}}-1}p+(1-p)^{N_{\mathrm{CN}}}=(1-p)^{N_{\mathrm{CN}}-1}, & n=N_{\mathrm{CN.}}\end{cases}$$

Given the distribution, it is straightforward to calculate other relevant quantities, such as moments. In particular, the mean number of checks is equal to

$$\bar{R}(N_{\mathrm{CN}},p)\triangleq \mathrm{E}(R)=\sum_{n=1}^{N_{\mathrm{CN}}-1}np(1-p)^{n-1}+N_{\mathrm{CN}}(1-p)^{N_{\mathrm{CN}}-1}=\frac{1-(1-p)^{N_{\mathrm{CN}}}}{p}. \qquad (1)$$

Always, $1\le \bar{R}(N_{\mathrm{CN}},p)\le N_{\mathrm{CN}}$. In accordance with intuition, the mean number of checks is a decreasing function of p, with $\lim_{p\to 1}\bar{R}(N_{\mathrm{CN}},p)=1$ and $\lim_{p\to 0}\bar{R}(N_{\mathrm{CN}},p)=N_{\mathrm{CN}}$. Equation (1) can be modified accordingly for networks with different suitability probability at the expense of

somewhat more complicated expressions for the results and some extra notation to express the asymmetries.

5.3.4 Delay components

The steps depicted in the MSC may be used as a guide to calculate the overall delay from the handover trigger to the completion of the VHO preparation phase. The overall delay can be calculated by keeping full account of message exchanges between the entities and summing up all the individual delays. Each step of the handover preparation process introduces a delay associated with the transmission, processing, or queueing of signaling messages. These sources of delay are discussed further in the following.

5.3.4.1 Transmission delays

These occur when transmitting messages over communication links. Transmission delays depend on the size of the signaling message and on the bandwidth of each link. Given a link bandwidth B_L and a packet length P, the wired link transmission delay is

$$D_L = \frac{P}{B_L}. \qquad (2)$$

Considering the wireless links and assuming that the available wireless link bandwidth BW_{WL} is fairly shared between the MNs served by the specific RAN we have the following: Assuming that the RAN serves N_{MN} MNs, each MN is allocated a bandwidth equal to B_{WL} / N_{MN}, therefore, the wireless link transmission delay will be equal to

$$D_{WL} = N_{MN} P / B_{WL}. \quad (3)$$

Also, D_{WL} can be adjusted for other types of opportunistic scheduling to different types of networks. Therefore, as the number of MNs served by a specific RAN increases, the derived BW for each MN decreases, having an impact on wireless link transmission delay.

5.3.4.2 Processing delays

Delays of this kind occur during the processing of a received message towards preparing a corresponding reply message, for example when querying the status of a CN to determine the availability of resources therein. In order to model the delays linked to the processing of a message, we consider the related workload of a procedure L, processed in a server with speed F. The corresponding processing delay is

$$D = L/F. \quad (4)$$

The characteristics of the probability distribution for D are inherited directly from those of the workload.

5.3.4.3 Waiting delays due to queueing

Congestion due to message queuing is the third component of the overall delay. The points where the signaling messages are being propagated or processed constitute potential congestion points, as messages are served in a First In First Out (FIFO) manner. These congestion points may be observed (depending on the architectural approach used) at the RANs and/or the CSs involved.

The calculation of delay components due to queueing requires the specification of the rates with which the various messages arrive at the queues. The whole VHO process begins with the handover trigger, which causes a sequence of further messages of all other signaling

message types. Due to the causality in the MSC there is correlation between the handover trigger and the succession of other messages. However, from a macroscopic network wide view, across a large number of MNs, the aggregate overall messages that are passing through a particular managing entity do appear to occur in a random and uncorrelated fashion, given that the timing characteristics of each sequence are independent from that of other sequences, thus they also follow a Poisson pattern, despite the deterministic association between messages.

Therefore, it can be regarded that all messages occur according to a Poisson process depending on the original rate of handover triggers per MN, in accordance with other parameters of the environment of the architecture.

More specifically, assuming λ handover triggers/sec per MN, in view of a homogeneous setup with a number of MNs per SN equal to N_{MN}, the overall rate of handover triggers per SN is

$$\lambda_{\text{triggers}}^{\text{SN}} = \lambda N_{\text{MN}}. \quad (5)$$

According to the MSC, each trigger corresponds to one query in order to acquire the list of CNs, so the rate of queries originating from the same SN would be equal to the value given in (5).

Each query is followed by a response with the same rate and a variable number of checks considering attempts to find a suitable handover target, querying the suitability of networks that appear in the list. As the MSC of Figure 9 indicates, a single trigger corresponds to a variable number of checks. This number of checks is random, independently and identically distributed between MNs. The mean value of checks $\bar{R}$, is computed later on based on

network parameters. Since the handover trigger rate per MN is λ, the overall rate of messages originating from a single RAN corresponds to

$$\lambda_{checks}^{SN} = \lambda N_{MN} \overline{R}. \quad (6)$$

Since the arrival rate of messages follows the Poisson process the queues mentioned can be modeled as multiclass M/G/1 queues, noting that the service time of each message queued depends on the type of the message. Messages of the first type (such as messages *a* and *c* in shaded area 1, and last message in shaded area 2, Figure 9) just require propagation through the wired or wireless links, corresponding to service times given by (2) or (3). Messages of the second type, (i.e.: *a* messages of shaded area 2, Figure 9) require processing in the relevant component, corresponding to service times given by (4).

For example, consider a RAN, which acts as a gateway managing signaling load both from MNs that serves (acting as a SN) and other network entities. The RAN queue serves request and response messages from its serving MNs that depend on (2) and (3), as well as, messages from other SNs related with the "lookup for resources" process that depend on (4).

For a multiclass M/G/1 queue, consider a general case with *K* classes of customers arriving with rates λ_k and having service requirements with means $E(S_k)$ and second moments $E(S_k^2)$, for $k = 1,..,K$, out of which variance can be derived. For example, considering mean and variance for a service time depending on (4): $E(S_k) = \frac{E(L_k)}{F_k}$ and accordingly $Var(S_k) = \frac{Var(L_k)}{F_k^2}$.

The class-specific traffic intensities are equal to $\rho_k = \lambda_k E(S_k)$, for a total load computed as the sum of the loads corresponding to each packet class $\rho = \sum_{k=1}^{K} \rho_k$. The queue is stable

exactly when $\rho < 1$. Although classes of customers have different requirements, they all experience the same waiting time (which can be defined as the mean waiting time experienced in a queue) of the same distribution [**122**]. The mean waiting delay $E(W)$ can be expressed as:

$$E(W) = \frac{1}{1-\rho}\sum_{k=1}^{K}\frac{\lambda_k}{2}E(S_k{}^2) = \frac{1}{1-\rho}\sum_{k=1}^{K}\rho_k E(S_k)\frac{1+Cv^2(S_k)}{2}. \quad (7)$$

Where, the coefficient of variation can be derived considering the mean and the variance $C_V^2(S_k) = \frac{Var(S_k)}{E(S_k)^2}$.

Accordingly, (7) can be specialized for priority-based queues at the expense of somewhat more complicated expressions and some extra notation to express the asymmetries.

The mean sojourn time for waiting plus service in the queue of class *k* is equal to:

$$E(Q_k) = E(W) + E(S_k), \text{ for } k = 1,...,K. \quad (8)$$

Therefore, calculations of (7) and (8) require the calculation of the arrival rates per class, derived from the equations (5), (6) and the two first moments of the service delay distributions per class (i.e. mean and variance), derived from the equations (2), (3), (4) (where the first and second order properties of *P* and *L* are involved).

We have already provided all the basic techniques and elements for the assessment of the overall delay of the VHO preparation phase, and now we can move on to the application of the presented methodology to two different architectural approaches, in order to compare them. In the interest of clarity and of keeping the presentation simple, the following developments assume a homogeneous setup, where all RANs serve the same number of MNs

and have the same wireless link characteristics and where links in the wired backbone are assumed to have the same capacity.

The elements of the modeling methodology can be put to use for the calculation of the mean delay associated with the entire VHO preparation phase: As a first step, the generic MSC of Figure 9 must be customized for the particular VHO architecture under study, providing detailed specifications for the signaling between the entities involved in phase 2, in accordance with the guidelines provided in the previous paragraphs. Queueing locations must also be identified at this point, together with an enumeration of the types of signaling messages handled by each queue.

Then, assignment of arrival rates and service times for each type of message proceeds as discussed in Chapter 6.2. The sojourn time at each queue is determined through appropriate applications of (8). Finally, the mean value of the overall latency is obtained by tracking a typical realization of the MSC and adding the individual delays due to transmission, processing and queueing associated with each step therein.

6. Performance Evaluation of an On-demand approach

6.1 On-demand approach architecture

In the following paragraphs, we introduce the On-demand Resource Information Gathering (ORIG) approach, which follows the directions of the major standards 3GPP ANDSF and IEEE 802.21, as described earlier. The main architectural entities of the network environment of Figure 8 are applicable, where the only CS involved is the IS, as depicted in Figure 10. The IS is considered to be centralized and responsible to keep the static information about the characteristics and services provided by the serving and neighboring networks, as it has been described by the ANDSF (i.e. ANDSF server) and IEEE 802.21 (i.e. MIIS) standards. The distinguishing feature of this approach is that the load information acquisition strategy is reactive (i.e. on-demand), interacting with each neighboring network, one by one, every time a handover is triggered.

The generic MSC of Figure 9 is further analyzed on the MSC of Figure 11. The first phase of the VHO preparation process is depicted in shaded area 1, corresponding to shaded area 1 of Figure 9, starting with the VHO trigger (now presented as an information request (IR) message) to the acquirement of the CNs list. Minor additions to the generic MSC include the IR response (IRR), which now is sent to the MN (see message d), in order to initiate the second phase, sending a query resources request (QR) message to the SN (see message e). The second phase is depicted in shaded area 2, corresponding to shaded area 2 of Figure 9 (where it was depicted in an abstract form), including the discovery of the suitable network target and the relevant query resources response (QRR) back to the MN. This process is now depicted in detail in Figure 11, showing the signaling required between the SN and the CNs

and the processing required at each CN to check about its current resources, involving a random number of checks until finding a network target with suitable resources. In the same figure it can be noted that a queue is depicted at each RAN, as they constitute congestion points that possibly involve queueing, as it is going to be discussed further.

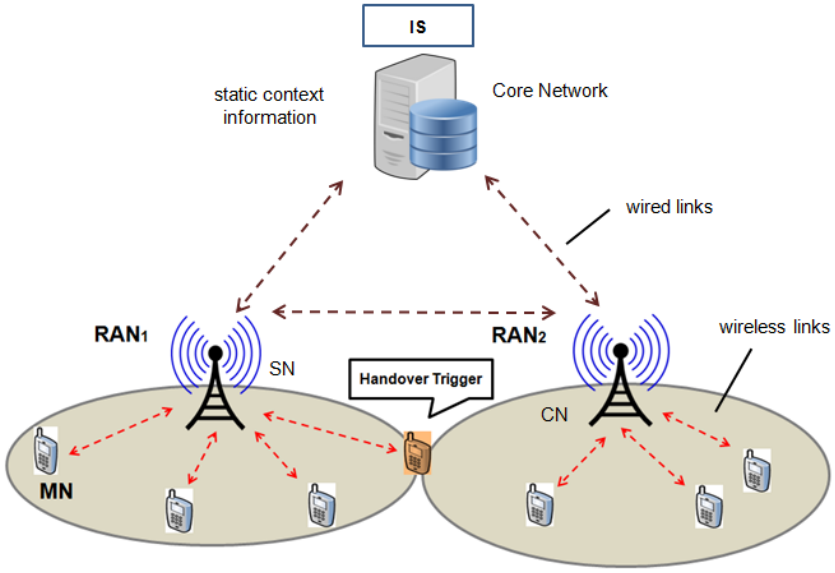

Figure 10. Main entities concerning the ORIG architectural approach.

6.2 Analysis of the On-demand Approach

We proceed further in putting the various time components together towards computing the overall delay for the VHO procedure as described above, which is based on the relevant application of the proposed generic modeling methodology. As depicted on the MSC of Figure 11, each RAN acts as a gateway managing a certain signaling load, because of its role as a SN, mediating between the MNs and the IS, and between the MNs and other RANs that act as CNs, as well. At the same time, each RAN acts as a CN for the MNs that are currently being served by other RANs. For these reasons, the gateways associated with the RANs are

modeled in as queues that serve all the incoming requests, responses and queries both from the network and from the MN sides.

It is also noted that queueing phenomena could have been considered at the IS, too. The reason for which this has not been pursued is that the IS is considered centralized, serving many more network entities beyond those whose performance is considered in the model. For a properly dimensioned IS in such a setting, traffic (or other parameter) changes related to the networks under examination would not have a significant impact to the magnitude of the waiting delay experienced at the IS queue. Consequently, this delay has been incorporated into the overall IS-related "processing delay" (of mean $E(D_{\mathrm{IS}})$).

Within the specific application of M/G/1 queues for modeling each RAN, three classes of customers are assumed, which are related to the three types of service times required by different signaling messages. The service times for each class can be computed as follows.

The class A messages are those directed to the wired links (towards the IS or the CNs) with service time S_A, as defined by (2), while, the class B messages are those directed to the wireless links (towards the MNs) with service time S_B, as defined by (3). The class C messages are the QRs received by a RAN that acts as a CN, requiring processing time equal to D_{Lookup}^{RAN} for each check, which can be further analyzed to the workload of the "lookup for resources" L_{Lookup} in relation to the processing speed of the RAN F_{RAN}, as defined by (4). In this case the processing time of D_{Lookup}^{RAN} is followed by a wired transmission time for returning the reply to the querying SN. Accordingly, S_C can be modeled as the sum of two independent random variables i.e., as $\frac{L_{Lookup}}{F_{RAN}} + D_L$. Therefore, the mean and the variance of

service time S_C are derived by the means or variances of the sum of two independent random variables, namely L_{Lookup} and D_L. For example, $\mathrm{Var}(S_A) = \mathrm{Var}(P)/B_L^2$ and $\mathrm{Var}(S_C) = \mathrm{Var}(L_{\mathrm{Lookup}})/F_{\mathrm{RAN}}^2 + \mathrm{Var}(P)/B_L^2$.

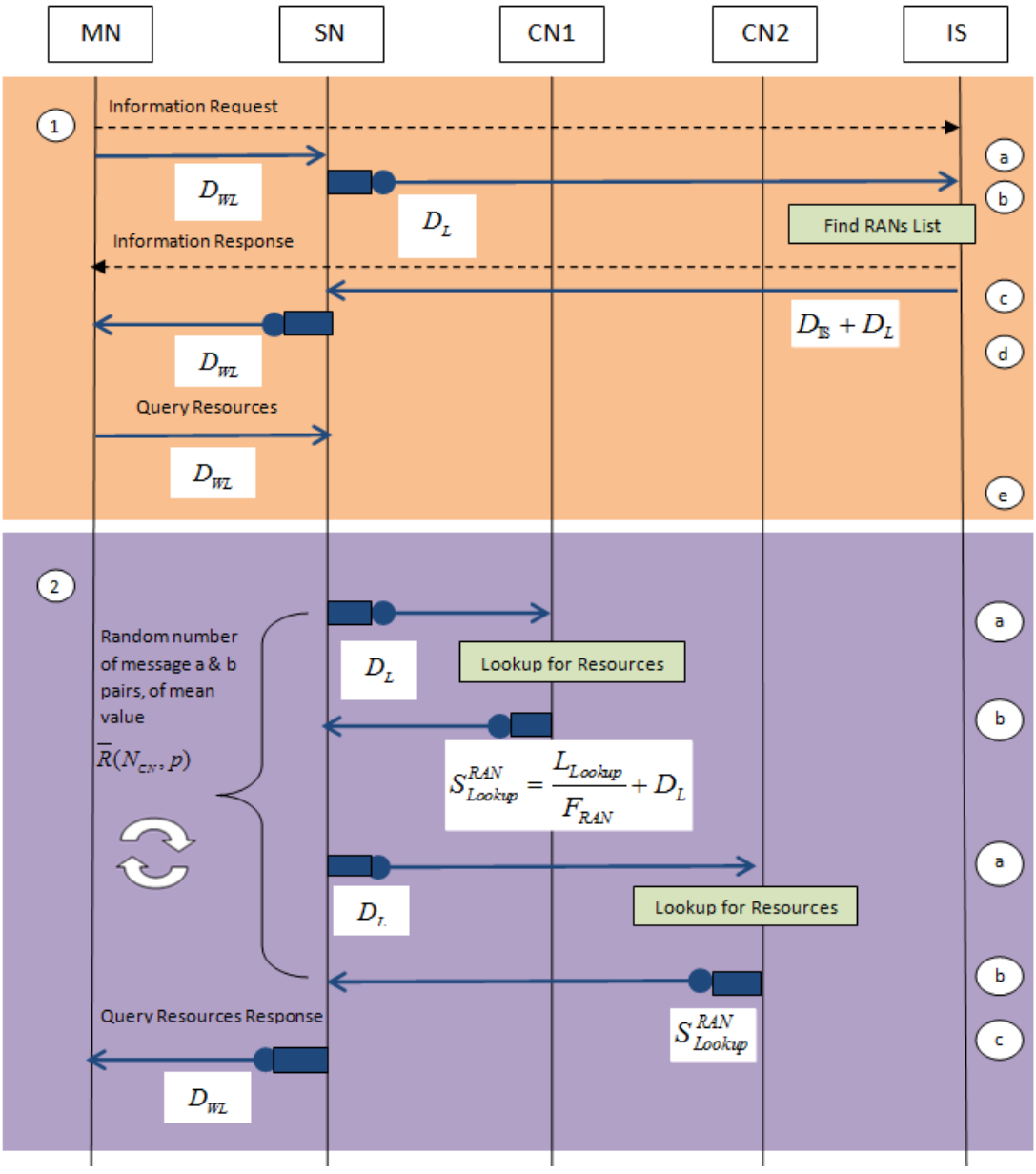

Figure 11. MSC for VHO preparation phase according to the ORIG approach.

The rate of IRs originating at a SN corresponds to $\lambda_{triggers}^{SN}$, described by (5). The overall rate of QR messages originating from a single RAN to all other CNs, considering all MNs served by this RAN, λ_{checks}^{SN}, has been described by (6), in accordance to the calculation of mean number of attempts per MN per trigger $\bar{R}$, as described by (1).

Now, we have to calculate the occurrence of the overall rate of QR messages that are received by a specific CN, considering that these messages have originated from all other neighboring RANs. A homogeneous setup is considered, where the number of neighboring RANs is denoted as N_{RANs}, while, the MNs that are being served by any given RAN and are being subject to handover are assumed to be able to consider all $N_{CN} = N_{RANs} - 1$ other RANs as candidates for the handover target.

When determining the distribution of the random number of checks, N_{CN} CNs are involved and thus $_R$ corresponds to $R(N_{CN}, p)$, where p signifies the probability of finding suitable available resources at a specific CN, as presented in the previously.

Consider a RAN that acts as a CN for its own MNs, but also as a CN for all other N_{CN} RANs. Thus, the RAN receives $\frac{\lambda_{checks}^{SN}}{N_{CN}}$ QR requests from each neighboring RAN, resulting to a rate of λ_{checks}^{SN}, considering the total number of all other RANs (which is equal to N_{CN} as described in (1)). The model can be readily extended, along the same principles, for addressing a heterogeneous setup.

The queue associated with each RAN receives messages that correspond to the three classes as mentioned before, according to the following rates (determined by inspection of Figure 11) as

$$\lambda_A^{ORIG} = \lambda_{triggers}^{SN} + \lambda_{checks}^{SN} = \lambda N_{MN}\left(1+\overline{R}(N_{CN}, p)\right), \quad (9)$$

$$\lambda_B^{ORIG} = 2\lambda_{triggers}^{SN} = 2\lambda N_{MN}, \quad (10)$$

$$\lambda_C^{ORIG} = \lambda_{checks}^{SN} = \lambda \mathrm{N}_{\mathrm{MN}}\,\overline{R}(N_{CN}, p). \quad (11)$$

The rate in (9) results from the occurrence of IR rate (5) and QR rate (6) of class A messages towards the IS and the CNs, accordingly, while rate in (10) results from IRR rate (5) and QRR rate (6) of class B messages towards the MNs. It is reminded that both rates in (9) and (10) involve messages sent by the RAN, acting as a SN. On the other hand, rate in (11) results from the occurrence of QR rate (5) of class C messages towards other RANs, sent by the specific RAN, acting as a CN.

Finally, we calculate the total mean delay for the VHO preparation phase, by accumulating the time components of the messaging sequence depicted in Figure 11, which involves the sum of delays spent inside the queues (i.e. waiting, processing and link delays), as well as, transmission delays over the links without a queue involved. To compute the mean waiting delay at a RAN queue $E(W_{RAN}^{ORIG})$ we have to make use of equation (7), computing the two first moments (i.e. mean and variance) of the service delay distributions per class S_A, S_B, S_C , and making use of the arrival rates computed in (9), (10) and (11). Considering the mean system delay for of each class in the queue, we use equation (7). It should be noted that delays relevant to the processing of QR requests spent at each RAN must be multiplied by the factor $\overline{R}(N_{CN}, p)$, to account for the multiple attempts involved. Mean transmission delays over the wired and the wireless links, denoted by $E(D_L)$ and $E(D_{WL})$, can been computed by (2) and (3) respectively and further, multiplied according to how many times they are involved, following the MSC of Figure 11.

After following the various steps of this calculation, the mean overall delay for the ORIG approach is

$$
\begin{aligned}
\mathrm{E}(\mathrm{D}_{\mathrm{ORIG}}^{\mathrm{TOTAL}}) = {} & \mathrm{E}(D_{\mathrm{WL}}) + \mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}) + \mathrm{E}(D_{\mathrm{L}}) + \mathrm{E}(D_{\mathrm{IS}}) + \mathrm{E}(D_{\mathrm{L}}) \\
& + \mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}) + \mathrm{E}(D_{\mathrm{WL}}) + \mathrm{E}(D_{\mathrm{WL}}) \\
& + \bar{R}(N_{\mathrm{CN}}, p)\left((\mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}) + \mathrm{E}(D_{\mathrm{L}})) + (\mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}) + \mathrm{E}(\frac{L_{\mathrm{Lookup}}}{F_{\mathrm{RAN}}}) + \mathrm{E}(D_{\mathrm{L}})) \right) \\
& + \mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}) + \mathrm{E}(D_{\mathrm{WL}}) \\
= {} & 4\,\mathrm{E}(D_{WL}) + \mathrm{E}(D_{\mathrm{IS}}) + 2(\bar{R}(N_{\mathrm{CN}}, p) + 1)\,\mathrm{E}(D_{L}) + \bar{R}(N_{\mathrm{CN}}, p)\frac{\mathrm{E}(L_{\mathrm{Lookup}})}{F_{\mathrm{RAN}}} \\
& + (2\bar{R}(N_{\mathrm{CN}}, p) + 3)\,\mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}}).
\end{aligned}
$$

(12)

The first equality in (12) includes separate terms for the parts of the latency associated with individual messages, also using parentheses for grouping together relevant subordinate terms. The terms are presented in the order the respective messages occur in the MSC of Figure 11. In particular, the parenthesized expression within the third line of (12) is equal to the mean latency associated with a single pair of messages a and b within phase 2. As explained earlier, a typical realization of the MSC includes a random number of such message pairs, to implement the required number of CN checks. On averaging, this introduces a factor equal to the expected number of checks, given in (1). The second equality in (12) expresses the result in a more compact form, grouping delay components further.

7. Performance Evaluation of a Proactive approach

7.1 Proactive Approach Architecture

We introduce the Proactive Resource Information Gathering (PRIG) scheme, which extends the standard VHO management frameworks IEEE 802.21 and 3GPP ANDSF. The distinguishing feature of this approach is that the load information acquisition strategy is proactive, introducing a local context repository that gathers load related context, periodically provided by the RANs. The proactively gathered contextual information would then be readily available to the requesting MNs, resulting potentially to a faster discovery of a network target with suitable available resources for the MN to handover to. In the following paragraphs, we illustrate the exact process of how resource-related information is made available, adapting the generic methodology of section 3. We demonstrate the relevant architectural entities involved and the related signaling, measuring all the relevant delays.

The distinguishing feature of the PRIG scheme is that the resource availability (which varies with the traffic load) is determined on the basis of information collected from the RANs proactively and periodically. Another CS, called "Dynamic Context Repository" (DCR), is introduced to manage the proactively gathered dynamic context. This arrangement simplifies the required signaling, by avoiding the need for individual queries to the CNs during the handover preparation process. Figure 12 revises the outline of the generic architecture in Figure 8, to explicitly indicate the inclusion of the DCR in the PRIG scheme. It is noted that static information, in particular the context required for determining the list of CNs, is still handled by the IS, as with the ORIG scheme.

While the IS is a centralized component of global scope, intended to serve a great number of RANs (comparable to all RANs in a whole country), the scope of the DCR in principle can range from global to local, depending on the number of RANs being monitored and served by it. However, in view of the dynamic nature of the context managed by DCR, fully centralized global deployments would be susceptible to complexity and scalability issues. Therefore, here we consider DCR deployments of local scope, in accordance with recent research directions [**119**], [**120**].

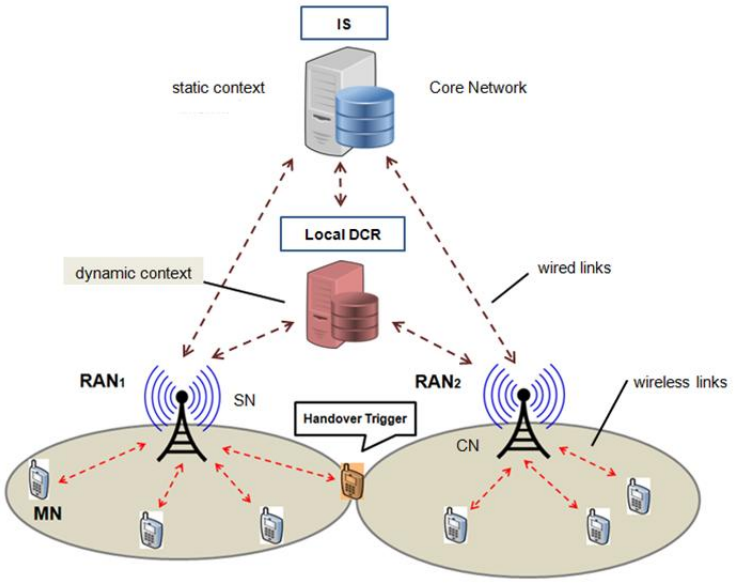

Figure 12. Main components of the PRIG architecture.

Specifically, it is assumed that the overall DCR functionality is provided via a number of DCR instances serving non-overlapping areas, each containing $N_{\mathrm{RAN}}^{\mathrm{total}}$ RANs in total. Given that our model considers N_{RAN} collocated RANs (i.e., RANs that could potentially be a CN for each other), the value of the ratio $N_{\mathrm{RAN}}^{\mathrm{total}} \big/ N_{\mathrm{RAN}} > 1$ quantifies the extent of each DCR instance's scope. Additionally, it is assumed that the DCR functions as an extension to the IS. This

implies that the RANs send/receive all context queries/replies to/from the IS, which interfaces with the (appropriate local instance of the) DCR when the query relates to dynamic context. Alternatively, the DCR functionality could be provided as a separate service, in which case dynamic context queries from the RANs would be sent directly to the DCR.

7.2 Analysis of the Proactive Approach

The customization of the generic MSC in Figure 9 for the PRIG case appears in Figure 13. The first two messages in phase 1 are as before (compare with the MSCs in Figure 9 and Figure 11), but now the IS sends the list of CNs directly to the appropriate instance of the DCR. This is in line with the assumption that the DCR functions as an extension of the IS. If the functionality was provided through a separate service, the list of CNs would be sent to the SN (message c of phase 1 within the MSCs of Figure 9 and Figure 11), which would subsequently query the DCR through an additional message. Moreover, by comparing with the MSC of the ORIG scheme (Figure 11), it can be seen that phase 2 is considerably simplified. The DCR employs the proactively collected dynamic context to per-form all required resource availability checks, without exchanging any signaling messages with individual CNs.

As already mentioned, the repetitive pattern of actions until determining the handover target is now reflected in the processing time required at the DCR. Note that the DCR sends the response (message a of phase 2 in Figure 13) to the IS, which subsequently forwards it to the MN via the SN (messages b and c). Again, this reflects the assumption that the DCR is an extension to the IS; if DCR functionality was provided through an additional service, message a would be sent to the SN instead. The messages depicted at the bottom of Figure 13 do not belong to the handover preparation process per se. These messages are sent proactively by all RANs in the domain of a DCR instance, to refresh regularly their resource availability status

and keep the dynamic context at the DCR instance up to date. This "resource information update" process occurs in the background, independently and in parallel with any handover preparation process that may be currently active. Each RAN is assumed to send update messages at a rate equal to λ^{updates}. Higher values of this rate correspond to more frequent updates, but also make the associated signaling load heavier.

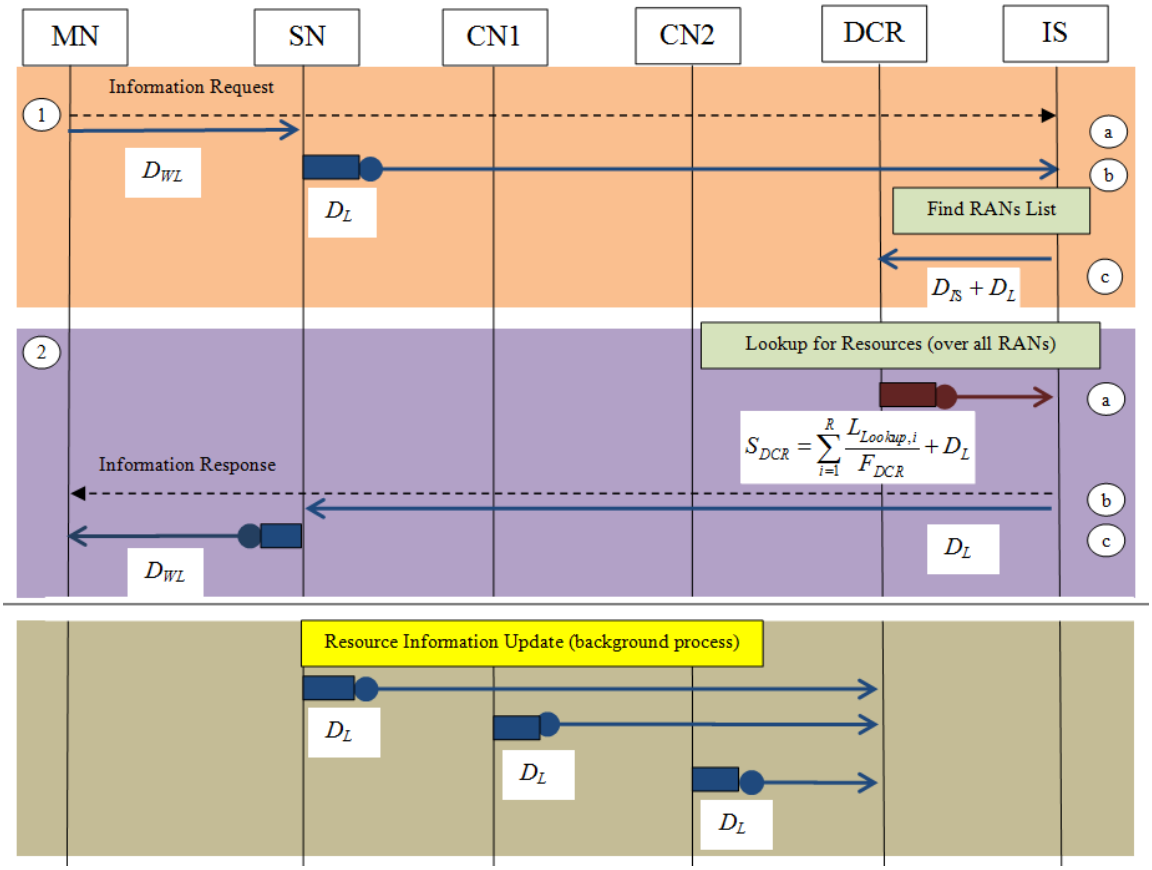

Figure 13. MSC for the VHO preparation phase according to the PRIG scheme.

As with the treatment of the ORIG case in Chapter 6, the MSC of Figure 13 indicates the latency associated with each message. It can be seen that, apart from queues at the RANs, now there is an additional queue at the DCR instance, to serve the responses to queries from the $N_{\text{RAN}}^{\text{total}}$ RANs in its domain. Since the DCR handles these queries without further RAN

involvement, the queue at each RAN handles a smaller load and involves messages belonging to classes A and B only, with service time characteristics as in the first and second line of (13). The respective message rates are equal to

$$\begin{aligned} \lambda_A^{\mathrm{PRIG}} &= \lambda_{\mathrm{triggers}}^{\mathrm{SN}} + \lambda^{\mathrm{updates}} = \lambda N_{\mathrm{MN}} + \lambda^{\mathrm{updates}}, \\ \lambda_B^{\mathrm{PRIG}} &= \lambda_{\mathrm{triggers}}^{\mathrm{SN}} = \lambda N_{\mathrm{MN}}. \end{aligned} \qquad (13)$$

Terms equal to $\lambda_{\mathrm{triggers}}^{\mathrm{SN}}$ in (13) correspond to occurrences of message b in phase 1 (for the first line) and message c in phase 2 (for the second), both these messages occurring once per handover trigger. Moreover, the expression for the rate of class A messages also accounts for the resource information updates sent proactively from the RAN to the DCR.

By comparing (13) with (10) it is seen that $\lambda_B^{\mathrm{PRIG}} = \lambda_B^{\mathrm{ORIG}} / 2$. This is because in the PRIG scheme the list of CNs is sent to the DCR instead of the MN, so fewer signaling messages are transmitted over the RAN's wireless link. Moreover, although PRIG introduces additional signaling for the proactive updates, this overhead does not become noticeable, because in typical settings $\lambda^{\mathrm{updates}} \ll \lambda_{\mathrm{checks}}^{\mathrm{SN}}$, so $\lambda_A^{\mathrm{PRIG}} < \lambda_A^{\mathrm{ORIG}}$ too. In view of the smaller load of class A and B messages and the absence of class C messages, the mean queueing time at a RAN queue $\mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{PRIG}})$ is typically much shorter than its counterpart $\mathrm{E}(W_{\mathrm{RAN}}^{\mathrm{ORIG}})$ in the ORIG scheme.

We now turn to the queue at the DCR, which handles a single class of messages, namely the responses to the resources availability queries. There is one such message for each handover trigger occurring in any RAN within the DCR instance's domain. Accordingly, the rate of messages at this queue is equal to

$$\lambda_{\mathrm{DCR}} = N_{\mathrm{RAN}}^{\mathrm{total}} \lambda_{\mathrm{triggers}}^{\mathrm{SN}} = \lambda N_{\mathrm{MN}} N_{\mathrm{RAN}}^{\mathrm{total}}. \qquad (14)$$

Clearly, as the scope of a DCR instance broadens, the load at the respective queue increases proportionally.

The service time of each message must reflect the whole sequence of CN checks performed until determining the handover target. Thus, a random number R of checks is involved, and the distribution of R is as discussed in Chapter 5. The processing of each CN check requires a random workload, distributed as the workload L_{Lookup} in the ORIG scheme. However, the whole sequence of checks now occurs at the DCR, employing the dynamic context maintained there. Thus, the total work-load at the DCR per message becomes

$$L_{\text{Lookup}}^{\text{tot}} = \sum_{i=1}^{R} L_{\text{Lookup},i}. \qquad (15)$$

All terms in the random sum are independent and identically distributed (iid). By factoring in the speed F_{DCR} of the processing facility at the DCR and including the additional time required for transmitting the processed message over the wired link towards the IS, the overall service time for each message at the DCR queue is seen to be equal to $S_{\text{DCR}} = L_{\text{Lookup}}^{\text{tot}} / F_{\text{DCR}} + D_{\text{L}}$, where the two terms are independently distributed. Consequently,

$$\begin{aligned} \mathrm{E}(S_{\text{DCR}}) &= \frac{\mathrm{E}(L_{\text{Lookup}}^{\text{tot}})}{F_{\text{DCR}}} + \mathrm{E}(D_{\text{L}}) \quad \text{and} \\ \mathrm{Var}(S_{\text{DCR}}) &= \frac{\mathrm{Var}(L_{\text{Lookup}}^{\text{tot}})}{F_{\text{DCR}}^2} + \mathrm{Var}(D_{\text{L}}). \end{aligned} \qquad (16)$$

It remains to calculate the mean and variance of the random sum in (15). This may be approached by first considering means and variances conditional on R and exploiting the fact that the terms in the summation are iid. Indeed, by the "law of total expectation" (see, e.g., equation (34.6) on p.448 of [**123**],

$$\begin{aligned} &\mathrm{E}(L^{\mathrm{tot}}_{\mathrm{Lookup}}) = \mathrm{E}(\mathrm{E}(L^{\mathrm{tot}}_{\mathrm{Lookup}} \mid R)) = \\ &= \mathrm{E}(R\,\mathrm{E}(L_{\mathrm{Lookup}})) = \bar{R}(N_{\mathrm{CN}}, p)\,\mathrm{E}(L_{\mathrm{Lookup}}), \end{aligned} \tag{17}$$

Similarly, by applying the "law of total variance" (see, e.g., problem 34.10(b) on p.456 of [**123**]), one obtains

$$\begin{aligned} \mathrm{Var}(L^{\mathrm{tot}}_{\mathrm{Lookup}}) &= \mathrm{E}(\mathrm{Var}(L^{\mathrm{tot}}_{\mathrm{Lookup}} \mid R)) + \mathrm{Var}(\mathrm{E}(L^{\mathrm{tot}}_{\mathrm{Lookup}} \mid R)) \\ &= \mathrm{E}(R\,\mathrm{Var}(L_{\mathrm{Lookup}})) + \mathrm{Var}(R\,\mathrm{E}(L_{\mathrm{Lookup}})) \\ &= \mathrm{E}(R)\,\mathrm{Var}(L_{\mathrm{Lookup}}) + \mathrm{E}(L_{\mathrm{Lookup}})^2\,\mathrm{Var}(R) \\ &= \bar{R}(N_{\mathrm{CN}}, p)\,\mathrm{Var}(L_{\mathrm{Lookup}}) \\ &\quad + \mathrm{E}(L_{\mathrm{Lookup}})^2 \;\; \bar{R}(N_{\mathrm{CN}}, p) \;\; 2(N_{\mathrm{CN}} + 1/p) - \bar{R}(N_{\mathrm{CN}}, p) - 1 \;\; - 2\,N_{\mathrm{CN}}/p \;\; , \end{aligned} \tag{18}$$

where the final equality follows from (1) and an explicit calculation of $\mathrm{Var}(R)$ according to the truncated geometric form of the distribution R.

By employing the service time characteristics in (16), (17) and (18), and the rate of messages in (14), the mean queueing time $\mathrm{E}(W_{\mathrm{DCR}})$ at the DCR queue can be calculated through a direct application of (8). Finally, the mean value of the overall latency is obtained by tracking a typical realization of the MSC in Figure 5. In the PRIG scheme, all such realizations always involve the same number of steps and the final result becomes

$$\begin{aligned} \mathrm{E}(\mathrm{D}^{\mathrm{TOTAL}}_{\mathrm{PRIG}}) = \; & \mathrm{E}(D_{\mathrm{WL}}) + \;\; \mathrm{E}(W^{\mathrm{PRIG}}_{\mathrm{RAN}}) + \mathrm{E}(\mathrm{D_L}) \;\; + \;\; \mathrm{E}(D_{\mathrm{IS}}) + \mathrm{E}(\mathrm{D_L}) \\ & + \;\; \mathrm{E} \;\; W_{\mathrm{DCR}} \;\; + \mathrm{E}(S_{\mathrm{DCR}}) \;\; + \mathrm{E}(\mathrm{D_L}) + \;\; \mathrm{E}(W^{\mathrm{PRIG}}_{\mathrm{RAN}}) + \mathrm{E}(D_{\mathrm{WL}}) \\ = \; & 2\,\mathrm{E}(D_{\mathrm{WL}}) + \mathrm{E}(D_{\mathrm{IS}}) + 4\,\mathrm{E}(\mathrm{D_L}) + \bar{R}(N_{\mathrm{CN}}, p)\frac{\mathrm{E}(L_{\mathrm{Lookup}})}{F_{\mathrm{DCR}}} \\ & + 2\,\mathrm{E}(W^{\mathrm{PRIG}}_{\mathrm{RAN}}) + \mathrm{E} \;\; W_{\mathrm{DCR}} \;\; . \end{aligned} \tag{19}$$

Again, the first equality includes separate terms for the parts of the latency associated with individual messages and the terms are presented in the order the respective messages occur.

The second equality follows after substituting $\mathrm{E}(S_{\mathrm{DCR}})$ with its equivalent, using (16) and (17), and grouping delay components further.

Comparing the result in (19) with the one in (12), it is seen that the PRIG scheme requires fewer transmissions of signaling messages over wired or wireless links than ORIG. In particular, only half of the significantly more expensive wireless transmissions are involved. Moreover, in PRIG there are fewer times when messages wait at RAN queues and, as mentioned earlier, the mean waiting time at each such queue is shorter than in ORIG.

However, PRIG introduces another queue at the DCR, which may handle messages from many RANs. Moreover, the processing of each message at this queue is complex, involving resource availability checks related to multiple CNs. Therefore, the computational capacity of a DCR instance (expressed through the corresponding processing speed F_{DCR}) must be properly parameterized, towards maintaining queue stability and avoiding excessive queueing times. This aspect is explored further through relevant evaluation results in the next Chapter.

8. Evaluation Results and Comparative Assessment of the different schemes

8.1 Evaluation Setup and Related Metrics

We focus on a case study that enables mobility in HetNets, including cellular and WiFi integration, which triggers current research interest [**16**]. We consider a 5G networks environment, where heterogeneous multi-RATs and multi-layer networks co-exist. The scenarios under consideration include mobility between small-cells (e.g. WiFi APs) that exist in the area of the same macro-cell (e.g. LTE BS), enabling MNs to handover to the appropriate small-cell that would provide them with adequate resources in order to keep the desired QoS.

Considering the topology under study, a number of small-cell RANs that reside in the coverage area of one macro-cell are depicted, forming clusters (i.e. neighborhoods), as presented in Figure 14. Each small-cell RAN constitutes a potential CN for the rest of the neighboring small-cell RANs in the cluster. Accordingly, each RAN acts both as a SN for a number of MNs, and as a CN for the MNs that are currently served by other neighboring RANs. The IS, as described by the 3GPP ANDSF and the IEEE 802.21 standards (defined as the ANDSF and the MIIS server, accordingly) is also included in the topology under study, which is responsible to keep static contextual information about the RANs, and provide the CNs list.

The models developed for the NS2 simulation experiments are completely scalable, accommodating a variable number of MNs and RANs. The RANs have been modeled as multiclass M/G/1 queues, where packets were enqueued and dequeued, in order to investigate

the congestion overhead. Note that the only approximation used in the analytical models is that the mean waiting delay $E(W)$ on all queues was calculated considering that all packet arrivals on the M/G/1 queue follow the Poisson distribution. On the other hand, in the simulation models, only the initial arrival packet rates (i.e. the VHO trigger rate λ generated by each MN) follow the Poisson distribution, resulting to smoother subsequent packet arrival rates, which is closer to reality.

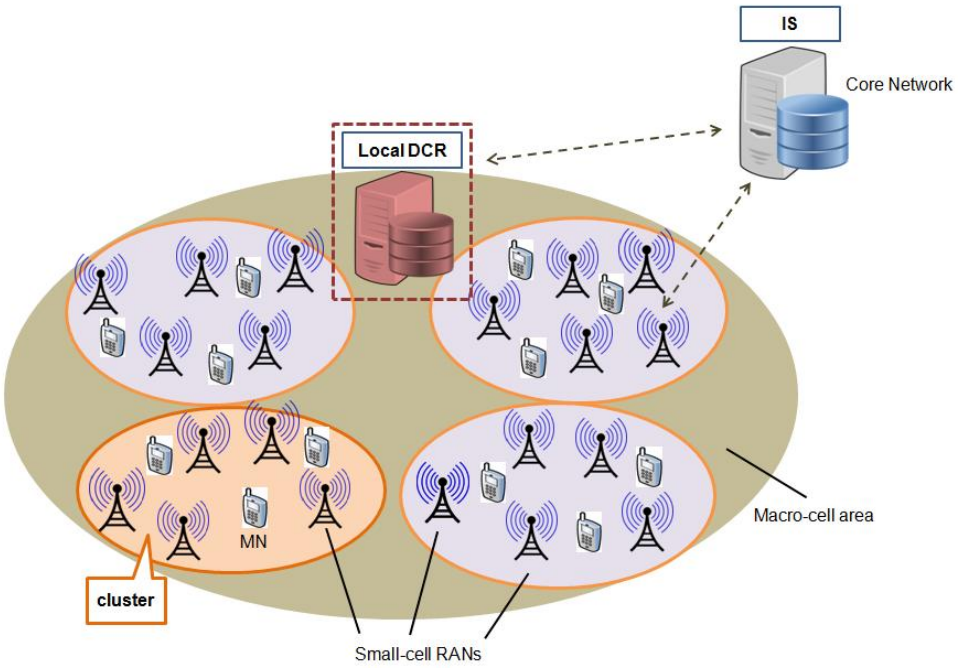

Figure 14. Topology of the evaluation setup.

The parameters used for the simulations and the analytical models are depicted in Table VI. Furthermore, in order to investigate the differences between diverse network topologies, we chose to present two scenarios. The size of the macro-cell area is taken equal to 3km2 (corresponding to a radius of approximately 1km), evenly divided into $N_{\text{cluster}} = 20$ clusters. Assuming an urban user density of 1000 users/km2 [**124**], there are 3000 MNs in the macro-cell area, evenly distributed therein. Using these parameters, we consider two deployment

scenarios: a scenario with a total number of $N_{\mathrm{RAN}}^{\mathrm{total}} = 100$ small-cells in the macro-cell area and a denser scenario with $N_{\mathrm{RAN}}^{\mathrm{total}} = 200$ small cells. In both cases, the number of small cells (neighboring RANs) per cluster is $N_{\mathrm{RAN}} = N_{\mathrm{RAN}}^{\mathrm{total}} / N_{\mathrm{cluster}}$ and all RANs except the SN are included in the list of CNs when a handover is triggered, i.e., $N_{\mathrm{CN}} = N_{\mathrm{RAN}} - 1$. Thus, in the second scenario there are twice as many small cells as in the first scenario (both in the macro-cell area and per cluster), each small cell serves half the number of MNs and each MN subject to handover must consider a greater number of CNs.

Furthermore, we have considered the following link and processing characteristics, also depicted in Table IV. More precisely, the wired link bandwidth can practically be regarded as a constant value, as it does not entail considerable variability.

Considering wireless link bandwidth, even in case of fast fading, which actually entails variability, variations happen so fast that practically the packet is transmitted with an average bandwidth value, which can be regarded as constant. Therefore, we use deterministic values for wired and wireless link bandwidth (corresponding to BW_L and BW_{WL}, respectively), according to typical WiFi values. Given the deterministic bandwidths, transmission delays over the wired and the wireless links (corresponding to D_L and D_{WL}, respectively) depend on the characteristics of the packet. Accordingly, we assume deterministic packet length P corresponding to deterministic transmission delays over the wired and the wireless links.

The processing times for checks for resources at the RANs have been considered as either deterministic or exponential, depending on the characteristics of the workload for one check L_{Lookup}, measured in work units, while for processing speeds F_{RAN} we have considered

deterministic work units/sec. The delay corresponding to the IS $D^{IS}_{RANsList}$ has also been considered to have deterministic values.

TABLE VI. Parameters used in scenario A and scenario B.

Parameter	**Value**	
	Scenario A	**Scenario B**
Macro-cell coverage area: $S_{macro-cell}$	$\pi A^2=3\times10^6$ m^2 (A=1000m)	
Number of MNs per macro-cell area (Urban: 1000 users/km^2) :	3000	
Number of small-cell RANs per macro-cell area: N^{total}_{RANs}	100	200
Number of clusters in a macro-cell area: $N_{clusters}$	20	
Number of neighboring RANs (CNs) in each cluster: N_{RANs}	5	10
Number of MNs per RAN: N_{MN}	30	15
Rate of VHO triggers per MN: λ (triggers/sec)	0.01	
Wired Link Bandwidth: $\mathrm{BW_L}$ (Mbps)	1000	
Wireless Link Bandwidth: $\mathrm{BW_{WL}}$ (Mbps)	100	
Mean message length $\mathrm{E}(P)$ (bits)	12000 (1500×8)	
Mean processing time at the IS $\mathrm{E}(D_{\mathrm{IS}})$ (sec)	0.01	
Mean workload per CN check $\mathrm{E}(L_{\mathrm{Lookup}})$ (work units)	1	
DCR Processing Speed F_{DCR} (work units/sec)	In multiples of F_{RAN} (variable)	
RAN Processing Speed: F_{RAN} (work units/sec)	100	
CN's resources suitability probability: p	[0.1, 1.0]	

The models developed for the simulation experiments are completely scalable, accommodating a variable number of MNs and RANs. The RANs and the DCR have been modeled as multiclass M/G/1 queues, where packets were enqueued and dequeued, in order to investigate the congestion overhead. Note that the only approximation used in the analytical models is that the mean waiting delay on all queues was calculated considering that all packet arrivals on the M/G/1 queue follow the Poisson distribution. On the other hand, in the simulation models, only the initial arrival packet rates (i.e. the VHO trigger rate generated by each MN) follow the Poisson distribution, resulting to smoother subsequent packet arrival rates, which is closer to reality.

8.2 Cost-Benefit Analysis Metrics

In order to address a cost – benefit analysis of each approach, we introduce a delay efficiency metric, defined as the (percentage-wise) improvement ratio of the difference between the PRIG and the ORIG mean overall delay, as $\frac{\mathrm{E}(\mathrm{D}_{\mathrm{ORIG}}^{\mathrm{TOTAL}}) - \mathrm{E}(\mathrm{D}_{\mathrm{PRIG}}^{\mathrm{TOTAL}})}{\mathrm{E}(\mathrm{D}_{\mathrm{ORIG}}^{\mathrm{TOTAL}})}\%$.

Furthermore, the computational resources expenditure related to each scheme has been considered. Specifically, computational resources expenditure can be translated in energy consumption expenditure, which is among the most important factors in the overall capital and operational expenditure of network operators [**125**], [**126**].

According to the ORIG scheme, the mean processing load that results to a single RAN, according to the class-specific traffic intensities, considering the processing of checks for resources (i.e. involving the class C messages) is $\rho_C^{\mathrm{RAN}} = \lambda_{\mathrm{checks}}^{\mathrm{SN}} \frac{\mathrm{E}(L_{\mathrm{Lookup}})}{F_{\mathrm{RAN}}} = \lambda N_{\mathrm{MN}} \bar{R}(N_{\mathrm{RAN}} - 1, p) \frac{\mathrm{E}(L_{\mathrm{Lookup}})}{F_{\mathrm{RAN}}}$.

Accordingly, considering the PRIG scheme, the mean processing load that results to a DCR, which involves the processing of checks for resources is

$$\rho^{\mathrm{DCR}} = \lambda_{\mathrm{DCR}} \frac{\mathrm{E}(\sum_{i=1}^{R} L_{\mathrm{Lookup},i})}{F_{\mathrm{DCR}}} = \lambda N_{\mathrm{MN}} N_{\mathrm{RAN}}^{total} \bar{R}(N_{\mathrm{RAN}} - 1, p) \frac{E(L_{\mathrm{Lookup}})}{F_{\mathrm{DCR}}}.$$

Apparently, the resulting load at the DCR is augmented by a factor of $N_{\mathrm{RAN}}^{\mathrm{total}}$ under the PRIG scheme, in relation to a single RAN, under the ORIG scheme. This means that the DCR is required to process proportionally $N_{\mathrm{RAN}}^{\mathrm{total}}$ more checks for resources than a single RAN. Therefore, theoretically, the required processing speed at the DCR should be $F_{\mathrm{DCR}}^{\mathrm{theory}} = N_{\mathrm{RAN}}^{\mathrm{total}} F_{\mathrm{RAN}}$.

However, practically, a much smaller processing speed is sufficient for the DCR, which varies according to each experiment, as it is going to be presented in the following. The applied value of processing speed at the DCR F_{DCR} can be measured as a multiple of the processing speed at a single RAN, as $F_{\mathrm{DCR}} = V \cdot F_{\mathrm{RAN}}$, with *V* defined as a processing speed ratio.

We introduce a processing efficiency metric, defined as the percentage-wise relative (to the theoretical value) reduction of the processing speed (i.e. processing saving) at the DCR which is equal to $\frac{F_{\mathrm{DCR}}^{\mathrm{theory}} - F_{\mathrm{DCR}}}{F_{\mathrm{DCR}}^{\mathrm{theory}}}\% = \frac{N_{\mathrm{RAN}}^{\mathrm{total}} - V}{N_{\mathrm{RAN}}^{\mathrm{total}}}\%$.

Processing efficiency is regarded in view of the fact that under the PRIG scheme the processing overhead of the resource checks among serving and candidate RANs is now transferred to the DCR, releasing the relevant processing resources of the respective RANs,

towards the goal for greener architectures [**127**], and lower operating expenditure (OPEX) and capital expenditure (CAPEX) costs.

8.3 Results on the Validation of the Analytical Model

In order to provide an elaborate view on the validation results considering the comparison of the analytical and the simulation models, we present the detailed results of the mean waiting delays that are observed, at the RAN queue, according to ORIG approach (Figure 15) under scenario A. This also enables the investigation of the impact of the distribution function of the checks for resources procedure at the RAN queue, which has been considered as either deterministic or exponential, as mentioned in the figures (while all the other types of relevant delays have been considered as deterministic). In general, considering the mean total handover preparation delay, the analytical results firmly coincide with the simulation results for both approaches, confirming the accuracy of the analytical models and that is the reason we provide the more elaborate view of Figure 15 and Figure 16, to point out any marginal differences.

It is interesting to point out that the difference between analytical and simulation models for the RAN queue, is almost negligible when $p = 1.0$, while it is slightly increasing as the candidate network resources availability probability is decreasing, as depicted in Figure 15. This captures the effect of the approximation used in the analytical models (as referred above) that all packet arrivals on the M/G/1 queue follow the Poisson distribution.

More precisely, under the ORIG approach, the end-to-end VHO preparation signaling involves a number of RAN queues, related to the probability of finding suitable available resources, which gradually causes higher divergence from the Poisson distribution at the

calculation of the mean RAN waiting delay, in comparison to the PRIG approach, as depicted in Figure 15 in comparison to Figure 16.

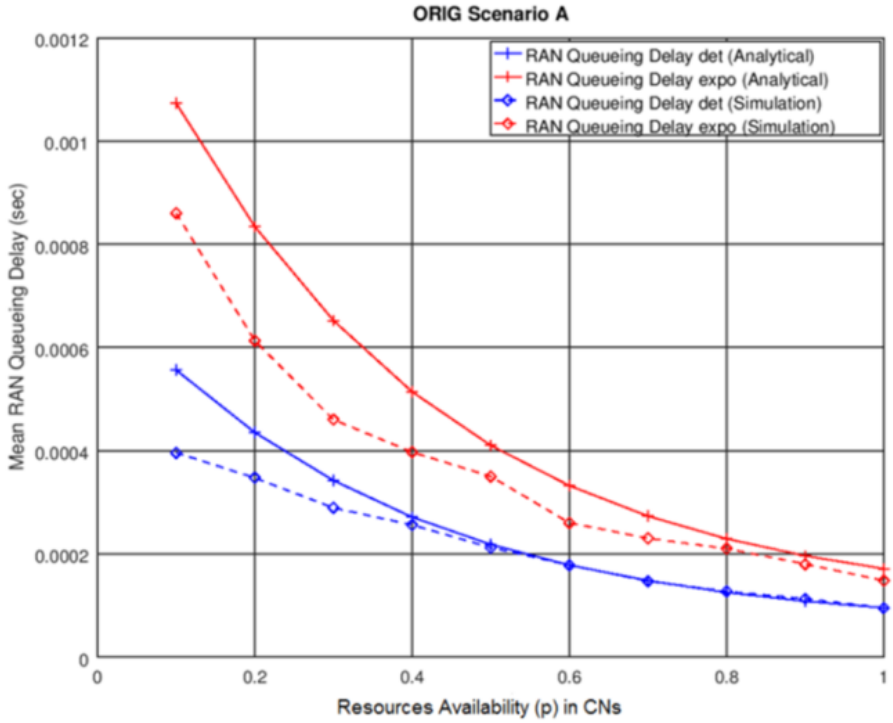

Figure 15. Mean RAN waiting delay for the ORIG scheme, under scenario A.

In general, it can be observed that the mean waiting delay resulting in the RAN queue (Figure 15) and the DCR queue (Figure 16), accordingly, is increasing as the probability of finding suitable available resources is decreasing, due to the increased signaling involved and the increased number of checks for resources, in order to find a suitable VHO target for the MN. It can also be noticed that in case the processing times for resource checks are considered as deterministic, the consequent queueing overhead is less than in the case that the processing times are considered as exponential. This is due to the fact that the mean waiting delay, as it is derived from equation (7), depends on the coefficient of variation (which is equal to 0 for deterministic delay, while it is equal to 1 for exponential delay).

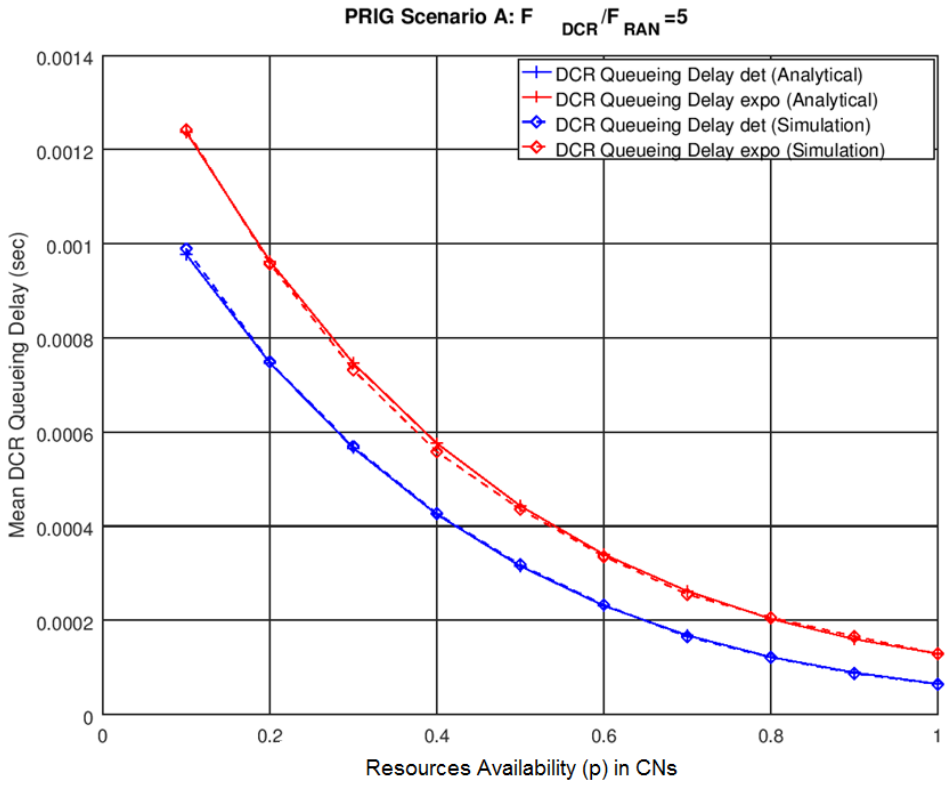

Figure 16. Mean RAN waiting delay for the PRIG scheme, under scenario A.

8.4 Results on the Performance of the On-demand Approach

The results on the performance of the ORIG scheme follows, under scenario A and scenario B, accordingly. Figure 17 (scenario A) and Figure 18 (scenario B) depict the mean end-to-end VHO preparation delay, presenting the amount of delay spent to the relevant (delay-contributing) components, according to p.

It can be observed that the "lookup for resources" component, involves a considerable amount of the overall delay, especially as the probability of finding suitable available resources is decreasing, and thus more CNs have to be checked until a suitable target network is found. Overall, the mean total delay of the ORIG scheme is increasing with a higher rate as p is decreasing, under scenario B.

More specifically, the number of CNs under scenario B is twice as much as under scenario A, affecting the term $\bar{R}(N_{RANs}-1, p)$ of equation (1) and thus involving more CN checks that burden the "lookup for resources" delay component, which is more apparent when $p < 0.5$. However, the number of the MNs per RAN affects the mean link delays that are much lower under scenario B than under scenario A, as the resulting wireless link bandwidth per MN is twice as much. In both scenarios, the congestion at the RAN queue is not significant, keeping the mean RAN waiting delay low.

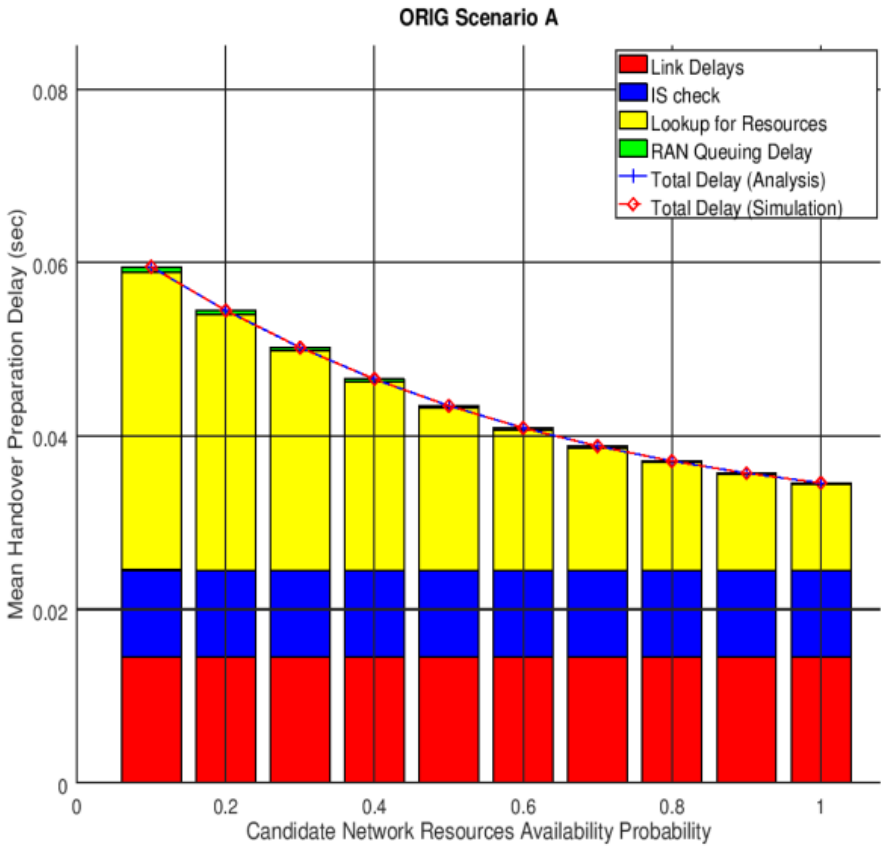

Figure 17. Mean VHO preparation delay, for the ORIG scheme under scenario A.

Overall, considering the applicability of the ORIG scheme in a next generation 5G networking environment, following the requirements of minimum end-to-end delay, it can be

seen that when p is high, the resulting end-to-end delay is low (up to 30 milliseconds) and thus acceptable. However, as p is decreasing and more queries are required to find a network target with suitable available resources, the end-to-end delay is increasing considerably, reaching up to 80 milliseconds, which would potentially cause degradation of QoS in time-critical applications.

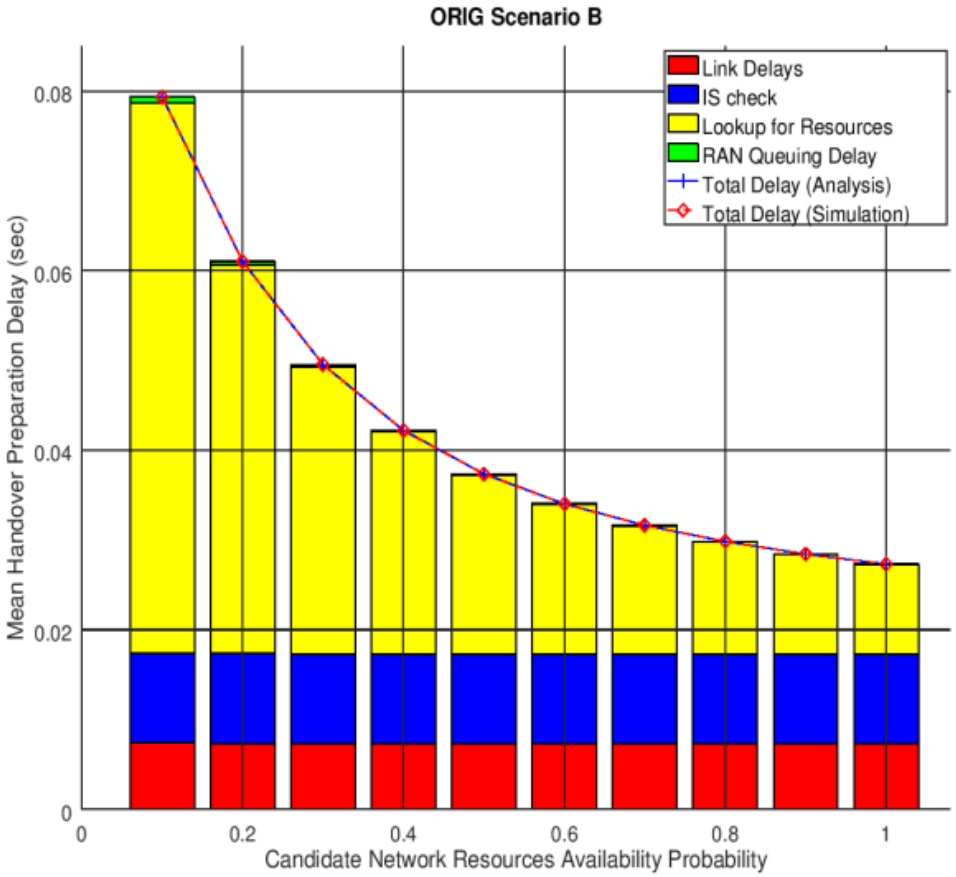

Figure 18. Mean VHO preparation delay, for the ORIG scheme under scenario B.

Therefore, the potential merits of the proactive context acquisition scheme, equipped with the capability of receiving and storing dynamic context proactively, have to be assessed and compared with the reactive on-demand approach.

8.5 Results on the Performance of the Proactive Approach and Comparative Assessment

Results on the performance of the PRIG scheme, as well as, the comparison of the two schemes (PRIG versus ORIG), follows, under scenario A and scenario B, respectively. Figure 19, depicts the mean end-to-end VHO preparation delay, presenting the amount of delay spent to the relevant (delay-contributing) components, according to p, under scenario A, for the PRIG scheme. In this case, the processing speed ratio $V = F_{\mathrm{DCR}} / F_{\mathrm{RAN}}$ is set to 5. A more elaborate view on the effect of the processing speed ratio is presented in the following paragraphs.

It is apparent that the PRIG scheme clearly outperforms the ORIG scheme, by inspecting Figure 19 in comparison to Figure 17, achieving delay efficiency from 44% to 58%, while only 5 times more processing speed is invested to the DCR in comparison to the case of a single RAN, which is not excessive resulting to 95% processing efficiency, following the fact that under this scenario the DCR serves a number of $N_{\mathrm{RAN}}^{\mathrm{total}} = 100$ RANs. Overall, the mean total delay of the ORIG scheme is increasing with a higher rate as p is decreasing compared to the PRIG scheme where the mean total delay is increasing with a much lower rate.

Specifically, it can be observed that the PRIG scheme involves almost half of mean link transmission delays in relation to the ORIG scheme, as it has already been derived by comparing the equation (19) in relation to equation (12). The presented transmission delays rely almost entirely on the wireless link transmission delays, as wired link delays are almost negligible. The mean delay spent due to the IS check is the same in both schemes, as it involves in any case one check at each VHO trigger.

Furthermore, it can be observed that according to the ORIG scheme, the "lookup for resources" component, involves a considerable time/component within the overall delay, especially as the probability of finding suitable available resources is decreasing, and thus more CNs have to be checked until a suitable target network is found. However, the congestion at the RAN queue is not significant, keeping the mean ORIG RAN waiting delay low under the specific scenario.

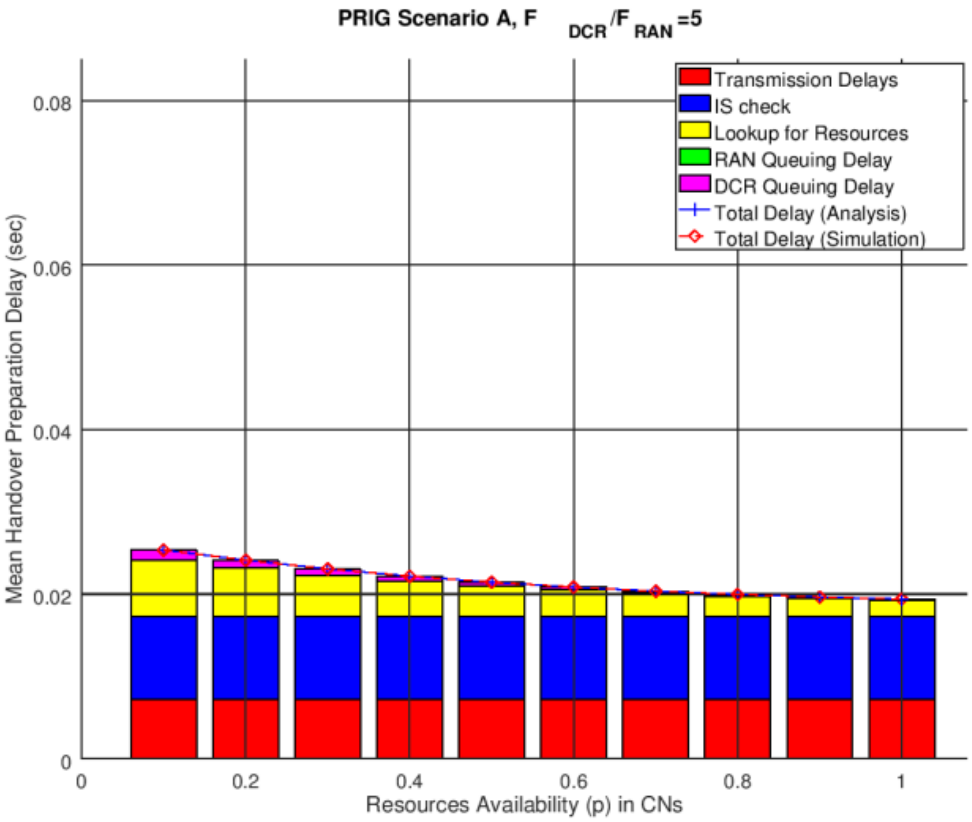

Figure 19. Mean VHO preparation delay, for the PRIG scheme under scenario A.

On the other hand, under the PRIG scheme, the benefits of the proactive resource information gathering strategy are obvious, as the "lookup for resources" component involves much lower amount of mean delay than under the ORIG scheme, considering the given

processing speed ratio. The mean PRIG RAN waiting delay is negligible, as no processing is involved at the RANs under this scheme. Nevertheless, the congestion at the DCR queue, represented by the mean DCR waiting delay, is considerable only when the probability of finding suitable available resources is low (i.e. $p < 0.3$), but still is not significant, under the specific DCR processing speed.

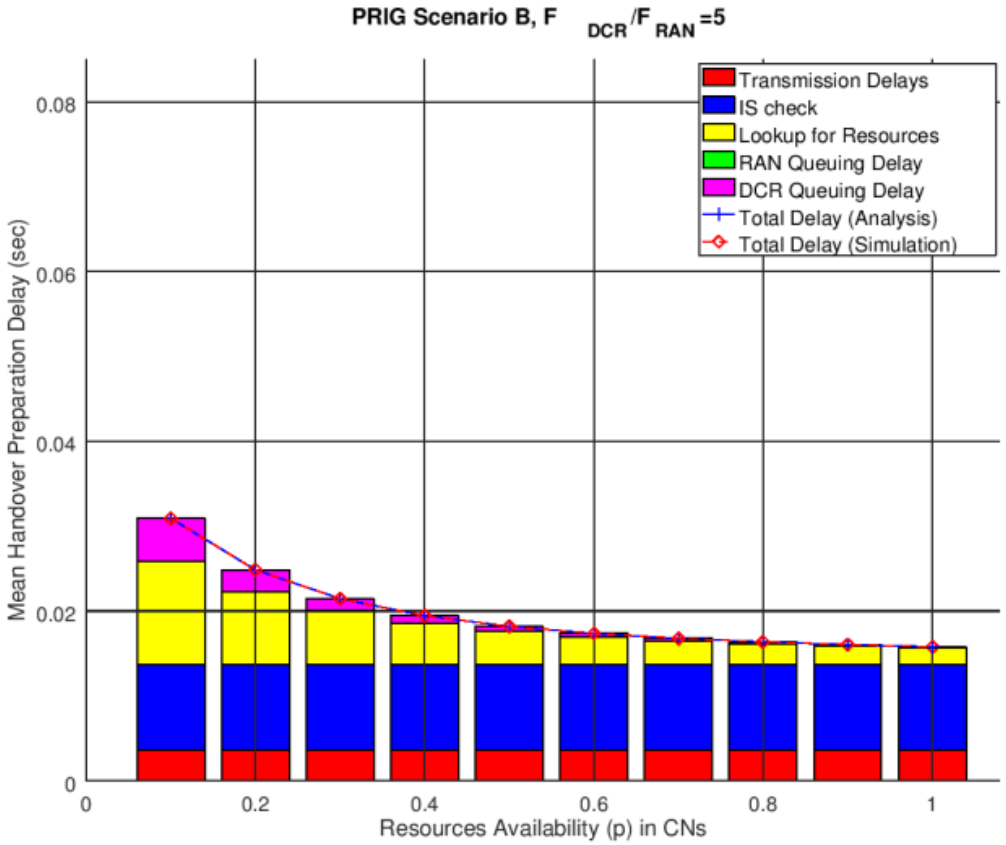

Figure 20. Mean VHO preparation delay, for the PRIG scheme under scenario B.

Proceeding to scenario B, Figure 20 depicts the mean VHO preparation delay, according to p, for the PRIG scheme, when processing speed ratio $V = F_{\mathrm{DCR}} / F_{\mathrm{RAN}}$ is set to 5, as with the aforementioned scenario, in order to have a fair comparison. Again, it is apparent that the PRIG scheme clearly outperforms the ORIG scheme, by inspecting Figure 20 in comparison

to Figure 18, achieving delay efficiency from 39% to 69%, with only 5 times more processing speed invested to the DCR in comparison to the case of a single RAN, resulting to 97.5% processing efficiency, following the fact that the DCR serves a number of $N_{\mathrm{RAN}}^{\mathrm{total}} = 200$ RANs under this scenario. Overall, by inspecting Figure 20 in comparison to Figure 19, the benefits of the PRIG scheme under this topology are even more obvious than those observed under the topology of scenario A.

More specifically, the number of CNs under scenario B is twice as much as under scenario A, affecting the term $\bar{R}(N_{\mathrm{CN}}, p)$ of equation (1) involving more CN checks that increase the "lookup for resources" delay component. For the same reason, the mean DCR waiting delay, according to the PRIG scheme, is more distinctive in cases of low probability of finding suitable available resources (i.e. $p < 0.3$), under this topology than under scenario A, but still it does not have a considerable effect to the overall delay values. Overall, the mean "lookup for resources" delay component under the ORIG scheme (see Figure 17 and Figure 18) is more heavily affected than under the PRIG scheme (see Figure 19 and Figure 20). Furthermore, the mean transmission delays are much lower under scenario B (see Figure 18 and Figure 20) than under scenario A (see Figure 17 and Figure 19), for both schemes, as the considered wireless link bandwidth per MN is doubled in scenario B.

For a more elaborate view on the effect of the processing speed ratio, Figure 21 and Figure 22, depict the delay efficiency of the PRIG scheme in relation to the ORIG scheme, under scenario A, and B, respectively, under different values of processing speed ratio V, for three different values of p, depicting good ($p = 0.8$), medium ($p = 0.5$) and weak ($p = 0.3$) probabilities of finding suitable available resources at each CN.

Considering scenario A (Figure 21), the minimum processing speed ratio V needed for the PRIG scheme to achieve the same mean total delay as that of the ORIG scheme is only 1.4 for $p = 0.8$, 1.6 for $p = 0.5$ and 1.8 for $p = 0.3$. It is also interesting that PRIG delay efficiency is increasing as p is decreasing, after a specific value of processing speed ratio (i.e. $F_{\mathrm{DCR}} / F_{\mathrm{RAN}} > 3$).

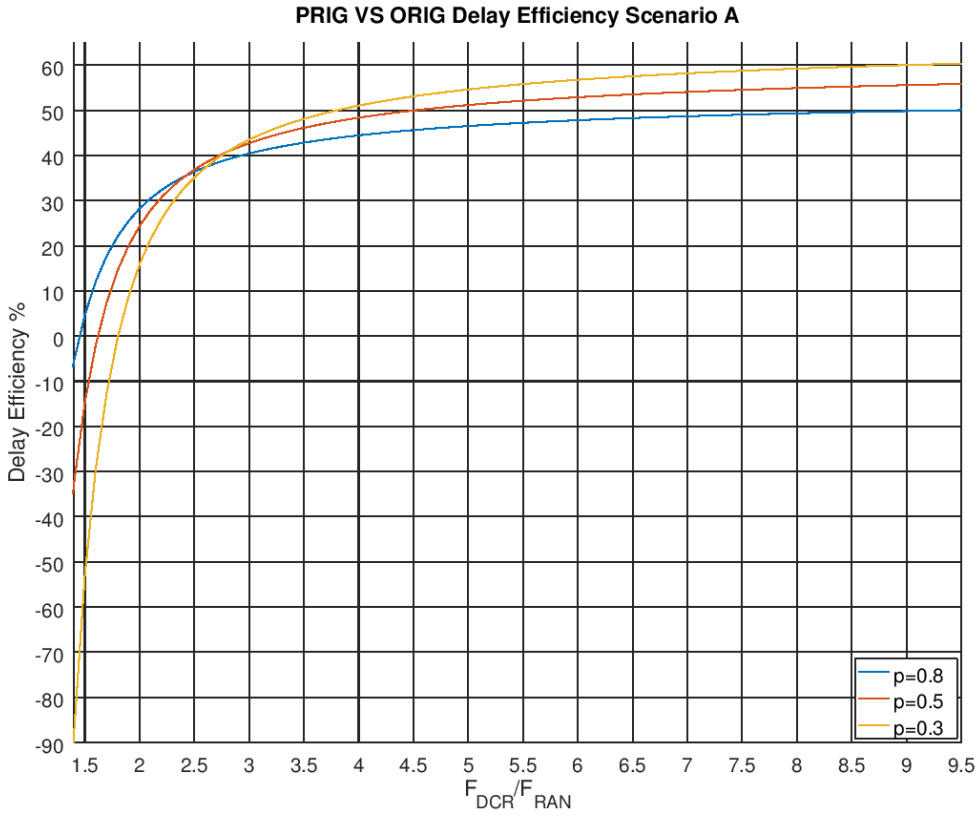

Figure 21. Delay Efficiency of PRIG in relation to ORIG scheme for scenario A.

This is due to the fact that the difference between ORIG and PRIG mean total delay is increasing with low p , rather than with high p , because the 'lookup for resources' delay component is increasing at a higher rate as p is decreasing, showing the benefits of the PRIG scheme (as it has already been shown/discussed in the preceding analysis). Accordingly, the

required processing speed ratio V needed for the PRIG scheme to achieve 50% delay efficiency is only 3.8 for $p = 0.3$, 4.6 for $p = 0.5$ and 9.5 for $p = 0.3$ which results to 96%, 95.4%, 90.5% processing efficiency, respectively (see Table VII).

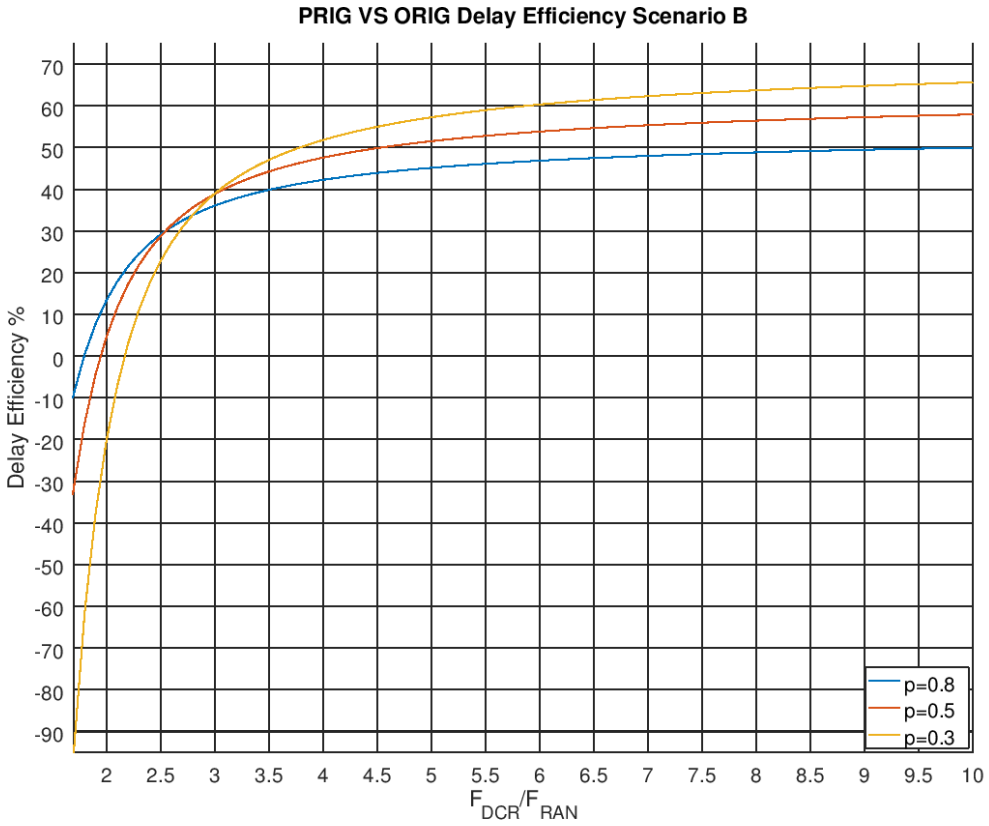

Figure 22. Delay Efficiency of PRIG in relation to ORIG scheme for scenario B.

In addition, it can be observed that the PRIG delay efficiency converges to a maximum value after a specific processing speed ratio V, because of the delay components that are not influenced by the processing speed (i.e. the transmission and the IS check delays).

Considering Scenario B (Figure 22), the minimum processing speed ratio V needed for the PRIG scheme to achieve the same the mean total delay as that of the ORIG scheme is only 1.7 for $p = 0.8$, 1.95 for $p = 0.5$ and 2.2 for $p = 0.3$. Again, it can be observed that delay

efficiency is increasing as p is decreasing, after a specific value of processing speed ratio (i.e. $F_{\mathrm{DCR}} / F_{\mathrm{RAN}} > 3.5$) under scenario B.

In this scenario, the required processing speed ratio needed for the PRIG scheme to achieve 50% delay efficiency in relation to the ORIG scheme is 3.8 for $p = 0.3$, 4.5 for $p = 0.5$ and 9.8 for $p = 0.8$ which results to 98.1%, 97.8%, 95.1% processing efficiency, respectively (see Table VII).

In addition, it can be observed that under the network topology of scenario B, the maximum achievable PRIG delay efficiency is up to 65% for $p = 0.3$, which is even higher in comparison to scenario A, which was up to 60%, as it was observed and discussed in Figure 21.

Table VII. Processing Speed Ratio (*V*) and Processing Efficiency, according to Delay Efficiency and *p*.

	Scenario A		**Scenario B**	
	V	*Processing Efficiency*	*V*	*Processing Efficiency*
Delay Efficiency 0%				
p=0.8	1.4	98.6%	1.7	99.1%
p=0.5	1.6	98.4%	1.95	99%
p=1.0	1.8	98.2%	2.6	98.7%
Delay Efficiency 50%				
p=0.8	9.5	90.5%	9.8	95.1%
p=0.5	4.6	95.4%	4.5	97.8%,
p=1.0	3.8	96%	3.8	98.1%

Concluding, the presented work aimed at providing a generic system model in order to evaluate the effect of different context-aware VHO management frameworks on the end-to-

end delay performance, which included the major architectural components, as well as, the significant methodological factors related to the analytical assessment of the overall delay. The model focused on capturing the effects of signaling in the VHO preparation phase, which is one of the most critical phases in the mobility management procedures.

The modeling methodology is comprehensive, yet able to produce closed form results, providing as much generality as possible in order to be versatile enough to adapt to different architectural VHO frameworks, extending the state-of-the-art approaches. The system model took into consideration the rate of the handover requests, the important network topological and availability characteristics (including the process of finding a suitable handover target), various sources of signaling overhead, the computational resources expenditure, and the congestion that results from all the aforementioned factors.

The generic system model and the proposed modeling methodology were used to compare the merits of the results of two different schemes that were based on proactive and reactive load information acquisition strategies, in order to select the appropriate network targets considering a HetNet environment of multiple different cellular layers (e.g. macro-small cell) and/or radio interfaces (e.g. 4G, 5G, WiFi), while catering to different service requirements.

The calculation of the mean end-to-end delay, as well as, the impact assessment of the computational resources' scaling on delay has been demonstrated in order to examine the feasibility of each approach considering efficient RAN selection in next generation networks. The principles of the proposed modeling methodology could be exploited for the subsequent study of additional architectural approaches that may be developed in the future.

More specifically, the reactive approach (i.e. ORIG), which checks the availability of resource-related information through interaction between the SN and each CN each time a

handover is triggered, was compared to a proactive approach (i.e. PRIG), which extends the standard VHO management frameworks, while, it presents a scalable and realistic architectural choice, by introducing a local dynamic context repository (DCR) that gathers load related context, proactively, periodically provided by the RANs, following the current research directions. The scenarios under consideration included mobility between small-cells within the coverage area of the same macro-cell, enabling MNs to handover from the macro-cell to the appropriate small-cell that would provide them with adequate resources in order to keep the desired QoS.

The simulation results confirmed the accuracy of the analytical models for both approaches. In addition, it was proven that the PRIG approach outperforms the ORIG approach, in both scenarios under study that are implemented under diverse network topologies, demonstrating considerable delay efficiency gains (up to 60% and 65%, accordingly), without excessive investments in computational resources for the DCR. Interestingly, it was proven that the PRIG scheme may present major processing efficiency gains (more than 90%), considering the overall processing resources expenditure, following the assumption that the processing resources of the various RANs are substituted by the local DCR, presenting potential energy consumption gains, towards the goal for greener architectures [**127**], and lower CAPEX and OPEX costs.

9. Concluding Remarks

9.1 Summary and Conclusions

In this study, the field of context-aware and autonomic VHO management was thoroughly explored. By employing concepts of ANM to VHO management, it became possible to shed new light to VHO operations from an ANM point of view, investigating the role of context-awareness and self-x capabilities, towards encompassing FI environments and the emerging 5G networks.

As a first step, basic concepts regarding context-awareness, cognition and autonomicity were reviewed, in **Chapter 2**. In the point of view taken, cognitive functions are considered as parts of ANM, characterizing a system aware of itself and its environment, self-governing its behavior to achieve specific goals. This view includes the notion of self-management. Subsequently, these concepts were employed in a classification and analysis of the components, processes and algorithms involved in autonomic handover management.

Ultimately, a new taxonomy of the relevant architectural components was introduced, considering the scenario of context-aware MNs that operate within a complex FI environment and self-manage their handover behavior. According to this taxonomy, the autonomic handover management procedure was organized into the phases of information collection, being linked to a knowledge base, followed by the handover decision making, which includes handover initiation and network selection processes and its corresponding algorithms, itself followed by the handover execution, which includes handover preparation and the related signaling. In this way, the VHO management complies with the autonomic management principles of monitor, analyze & plan, and execute functions.

Following this study's point of view, the standard media independent handover management frameworks were reviewed and correlated with the new autonomic framework. Relevant amendments and/or extensions to the standards from the literature were also reviewed, together with works addressing efficiency, performance evaluation and related modeling aspects.

As an additional contribution, this study highlighted a number of important autonomic features that may be leveraged to automatically make the autonomic handover management adaptive and to optimize its performance, towards the overall enhancement of the VHO operations, presented in **Chapter 3**. It was demonstrated that the combination of awareness, adaptivity & flexibility, learning and proactivity drive the system to performance improvements and enable the system to select the best choice among a set of available alternatives, advancing the system's overall self-optimization.

Robustness issues, related to the ability to achieve stable and efficient VHO decisions in the diverse and dynamic context of a FI environment, were also considered, filling a gap in the literature. A number of robustness-related issues and reviewed mechanisms to cope with them were identified, presenting also how robustness can be enhanced through the exploitation of autonomic features.

To demonstrate the applicability of the general concepts, a number of representative VHO management solutions with an autonomic orientation were reviewed taken from the literature, in **Chapter 4**. These solutions were analyzed and relevant characteristics were associated with the proposed classification and taxonomy. Furthermore, the solutions were compared in terms of the extent they incorporate and exploit the autonomic features identified, towards enhancing the effectiveness and efficiency of VHOs and achieving robustness. The principles employed in the

analysis and comparison of these particular solutions can be useful also for the future treatment of other VHO management solutions with an autonomic orientation.

In the course of this study, it was seen that, in order to provide seamless VHOs and enhanced decisions in a FI environment, there is need for information collection targeting parameters over multiple layers. This requires advanced context-awareness. A flexible and adaptive set of parameters is also key to a more effective support for the QoS requirements of running applications. Once the sophistication of the parameters set increases, it becomes important to ensure the robust operation of the system, even in unpredictable situations, by appropriately handling the diversity of parameters and criteria/rules, by providing resilience under context uncertainties and incompleteness and by including mechanisms to withstand marginal/borderline cases.

Also, the tradeoff between intelligence/sophistication and operational complexity was highlighted, which determines the achievable degree of self-management for the MN. While autonomic VHO management architectures should remain distributed enough to enable the MNs to make at least some decisions on their own, it may prove infeasible to implement sophisticated techniques solely on end-devices. In order for the system to keep its prompt reaction and maintain its potential for self-optimization, it may be advisable to introduce hierarchy in the VHO architecture and have specific time- and resource-demanding procedures linked with the cognition cycle (such as conflict resolution correlated to network-wide statistics and big data analysis, or learning algorithms) be coordinated by network entities higher up in the hierarchy, removing the burden from the MNs.

The second part of the study aimed at providing a generic system model in order to evaluate the effect of different context-aware VHO management frameworks on the end-to-end delay

performance, which included the major architectural components, as well as, the significant methodological factors related to the analytical assessment of the overall delay, presented in **Chapter 5**. The model focused on capturing the effects of signaling in the VHO preparation phase, which is one of the most critical phases in the mobility management procedures. The modeling methodology is comprehensive, yet able to produce closed form results, providing as much generality as possible in order to be versatile enough to adapt to different architectural VHO frameworks, extending the state-of-the-art approaches. The system model took into consideration the rate of the handover requests, the important network topological and availability characteristics (including the process of finding a suitable handover target), various sources of signaling overhead, the computational resources expenditure, and the congestion that results from all the aforementioned factors.

The generic system model and the proposed modeling methodology were used to compare the merits of the results of two different approaches that were based on reactive and proactive load information acquisition strategies, in order to select the appropriate network targets. The principles of the proposed modeling methodology could be exploited for the subsequent study of additional architectural approaches that may be developed in the future.

More specifically, the reactive on-demand approach (i.e. ORIG), was presented in **Chapter 6,** which checks the availability of resource-related information through interaction between the SN and each CN each time a handover is triggered. Furthermore, the proactive approach (i.e. PRIG), was presented in **Chapter 7,** which presents a scalable and realistic architectural choice, by introducing a dynamic context repository (DCR) that gathers load related context, proactively, periodically provided by the RANs, following the current research directions.

The calculation of the mean end-to-end delay, as well as, the impact assessment of the computational resources' scaling on delay has been demonstrated, in **Chapter 8**, in order to examine the feasibility of each approach considering efficient RAN selection in next generation networks. The scenarios under consideration included mobility between small-cells (e.g. WiFi APs) within the coverage area of the same macro-cell (i.e. LTE BS), enabling MNs to handover from the macro-cell to the appropriate small-cell that would provide them with adequate resources in order to keep the desired QoS.

The simulation results confirmed the accuracy of the analytical model for both approaches. It has been proven that the topology (i.e. the network density and the number of MNs per network), as well as, the context acquisition strategy, and especially the resource related information about the dynamic load of the networks, play an important role on the VHO performance metrics.

Overall, considering the applicability of the ORIG scheme in a next generation 5G networking environment, following the requirements of minimum end-to-end delay, it has been shown that the resulting end-to-end delay may reach up to 80 milliseconds, which would potentially cause degradation of QoS in time-critical applications. Therefore, the potential merits of an alternative context acquisition strategy, equipped with the capability of receiving and storing dynamic context proactively, have been assessed, following the proposed performance evaluation methodology that captures all important factors related with context-aware connectivity management in HetNets.

Specifically, it was shown that the PRIG scheme outperforms the ORIG scheme, in both scenarios under study that are implemented under diverse network topologies, demonstrating a considerable delay efficiency (up to 60% and 65%, accordingly), without excessive investments in computational resources for the DCR. In fact, major processing efficiency (more than 90%)

considering the overall processing resources expenditure is derived with the PRIG scheme, following the assumption that the processing resources of the various RANs are substituted by the local DCR, leading to potential energy consumption savings and lower OPEX and CAPEX costs.

9.2 Insights for Future Research

As this study comes to its completion, this section highlights and proposes future research directions, based on the outcomes of this doctoral thesis. Specifically, the following topics can be identified as main open lines of research, while insights related with the confrontation of the related issues are also provided.

- Current Issues considering Emerging Technologies related to Connectivity Management

Connectivity management in 5G networks is still an open issue, considering the augmenting densification of networks, the multitude of radio access technologies and the unplanned network layout, posing several challenges. Furthermore, 5G networks promote programmable and virtualized architectures, which can be integrated with Media Independent Handover (MIH) strategies in order to holistically improve the network performance [**128**]. Additionally, the strict latency requirements demand distributed network designs, shifting computing resources and storage at the edge of the network.

Specifically, emerging technological concepts, such as network virtualization and SDN, promote the design of more advanced and flexible network architectures [**129**]. SDN promotes the abstraction of the network logic from hardware implementation into software, proposing the separation of the data and control planes, and introducing a network controller

to coordinate the network's operations. In a similar direction, network virtualization can logically separate a single physical network infrastructure into multiple logical virtual networks, enabling customized support of application-specific services, and can lead to more efficient utilization of resources [**130**].

Several approaches attempted to integrate MIHs with the SDN architectural paradigm [**11**], [**128**], [**131**] and also with Network Function Virtualization (NFV) [**132**], implementing the network entities as Virtual Machines (VMs) [**133**], envisaging support for network slicing [**130**] referring to the existence of multiple, possibly isolated, service and network architectures to support different usage scenarios, in particular services hosted by different verticals.

Accordingly, emerging research concepts promote distribution of resources (i.e., compute, storage, and network resources) with multi-access edge computing (MEC) [**134**] minimizing the latency fluctuation, towards addressing the more stringent requirements of novel concepts from vertical industries, such as vehicular networks, considering not only decentralization for data, but also for control, proposing that physically distributed, yet logically centralized control plane, could be used to enhance performance.

Following this mentality, recent directions of ANDSF and Hotspot 2.0 integration, enable MNs to collect information from a local instance of ANDSF ("Local ANDSF" [**119**]) about the policies of the operator for accessing the various RANs in the area, as well as, dynamic information from Hotspot 2.0 protocols [**16**], as presented in Chapter 7. However, considering the state-of-art approaches, there exists a literature gap considering how to acquire and manage the plethora of context information, resulting from various entities – and how to optimize the tradeoff between context acquisition and signaling and processing

overheads and provide a feasible and realistic approach for efficient PoA selection in 5G HetNets, under the augmented network management requirements and challenges considering mobility between different RATs (i.e., including both 3GPP and non-3GPP access networks).

Novel, flexible and dynamically (re)-configurable network elements will be required to support these architectures and provide diverse and customizable services to dynamic traffic demands in frequency, space and time, while satisfying user QoS requirements. Towards the same direction, recent research directions suggest to move also the control closer to the edge in mobile networks, in order to overcome the limitations of centralization, proposing that physically distributed, yet logically centralized control plane could be used to enhance performance [**135**], [**136**]. Therefore, it becomes apparent that the application of the aforementioned emerging technologies on connectivity management demands an efficient network orchestration paradigm.

- Towards an Autonomic and Programmable Connectivity Management Architecture

To accommodate the aforementioned issues, we argue that a future wireless distributed connectivity management architecture should utilize principles of context-aware and ANM dictated by a consummated implementation and inter-dependence of SDN and NFV [**129**], [**137**] enabling the system to evolve and to adapt to changes, in terms of either business objectives or users requirements [**137**], following the outcomes of Chapters 2 to 4. Additionally, the methodology provided in Chapter 5, could be accordingly adapted to measure the performance of such future connectivity management architectures, towards optimizing the tradeoff between context acquisition and signaling and processing overheads and providing a feasible and realistic approach for efficient network selection in 5G networks and beyond.

More specifically, current research directions, promote context-aware strategies that demand feedback from the MNs. Following the outcomes resulted from the dissertation, related operations can be further optimized implementing self-x capabilities, which can be founded on the use of cooperative and cognitive radio strategies to reduce the mobile operator's maintenance and administration overhead, towards performance enhancement and minimization of the required energy consumption and delay overhead [**20**]. Towards these goals, ANM provides context-aware MNs that are able to self-manage their mobility behavior according to Policy-Based Management (PBM) principles.

Specifically, ANM introduces rules to formalize the description of operations of various network elements in response to changes in the environment, utilizing PBM, as highlighted in Chapter 2.1. Accordingly, as it was also highlighted in the dissertation, collection, modeling, reasoning, and distribution of context play a critical role in ANM, delivering to the system cognitive capabilities, which can be used in conjunction with machine learning and proactive techniques, exploiting context for responding faster and more efficiently to specific stimuli, and leading to overall system optimization.

Future autonomic network management architectures are likely to have a more flexible and customizable structure, towards adjusting dynamically the degree of self-management for each entity in the distributed architecture. Meanwhile, network elements higher up in the hierarchy should be able to enforce the appropriate policies on the MNs, in order to achieve global optimization goals. The relevant processes may combine ANM with SDN and network virtualization concepts, enabling a shift from device-driven management models to context-aware and QoS-aware management models, covering the market demand for more flexible and extensible network designs [**129**].

In such hybrid future autonomic architectures (following the architectural variances, highlighted in Chapter 2.2.4) some self-organization and self-optimization algorithms, mostly those related to tasks with local scope, would run locally on the MNs, while the tasks with wider scope (global network view) could be managed by a central managing authority (i.e. a network controller) on the network side. In particular, the global knowledge could provide enhanced information to MNs enabling dynamically optimized policies, towards achieving a local optimum in balance with the global optimum, according to an evolutionary process. In this way, the integration of autonomic MNs and autonomic network entities higher up in the hierarchy may lead to a VHO management solution featuring increased reliability and efficiency [**20**].

Accordingly, global awareness can be built by combining the self-awareness of each distributed entity, while MNs could utilize the global view to evaluate/verify and complement their own status. Towards the implementation of such global knowledge base, the strict latency requirements for mobility management in a 5G network environment, demand computing resources and storage at the edge of a network, utilizing edge, mobile edge, mobile cloud and fog computing concepts**.** Such distributed knowledge base paradigms could facilitate the MNs to store individually essential information about their mobility for later use, further promoting the concept of self-management. In this way, the goal of QoS-aware ubiquitous connectivity and efficient use/reuse of resources will be provided by a network architecture with advanced intelligence, capable of sensing its operational conditions and adapting its configuration accordingly.

- Context-Awareness and Autonomicity in conjunction with Cognitive Radios Addressing Multiple-Connectivity Opportunities

Spectrum is a limited resource and due to the fixed spectrum assignment policy and the rapid development of wireless networks, spectrum scarcity has significantly intensified. Some frequencies are heavily congested, e.g., the unlicensed spectrum, while others remain under-utilized. Cognitive radios allow for dynamic spectrum management, mitigating the aforementioned problems. Current Software Defined Radio (SDR) technology may act as an enabler, allowing a cognitive radio to configure dynamically the transmission parameters of a device, in accordance to the wireless environment in which it operates [**129**].

As a result, future research should focus on the interplay between SDN-SDR in distributed wireless networks operating in highly dynamic environments using NFV as a convergence substrate enabling the SDN-SDR interplay [**129**]. The objectives of the future network architectures should include reconfiguration flexibility, efficient use of the bandwidth, as well as, efficient and transparent Device-to-Device (D2D) communications, without interrupting the primary network operation.

As it has been highlighted in the course of this study, ANM provides context-aware MNs that are able to self-manage their mobility behavior according to policy-based management principles. Therefore, the combination of ANM and SDN-NFV-SDR can address the requirement of availability and resilience, especially in scenarios where the amount of devices or the network connectivity conditions are unsuitable for maintaining a frequent communication between mobile devices and centralized entities, promoting the accomplishment of novel distributed communication concepts, such as D2D [**129**].

Accordingly, we propose that the combination of ANM with SDN-SDR via virtual utility functions, would enable a cognitive and flexible framework to enable autonomic network management in cognitive radio network environments. Such framework following the cross-layer

design approach would address the objective of QoS-aware ubiquitous connectivity and efficient use of resources with flexible network management, utilizing multiple connectivity opportunities, in 5G networks and beyond.

10. Publications

PUBLICATIONS IN INTERNATIONAL JOURNALS

- **Adamantia Stamou**, Nikos Dimitriou, Kimon Kontovasilis, Symeon Papavassiliou, "Autonomic Handover Management for Heterogeneous Networks in a Future Internet Context: A Survey", *IEEE Communication Surveys and Tutorials*, 2019, DOI: 10.1109/COMST.2019.2916188.
- **Adamantia Stamou**, Nikos Dimitriou, Kimon Kontovasilis, Symeon Papavassiliou, "Context-Aware Handover Management for HetNets: Performance Evaluation Models and Comparative Assessment of Alternative Context Acquisition Strategies", *submitted for publication*, 2019.
- **Adamantia Stamou**, Grigorios Kakkavas, Konstantinos Tsitseklis, Vasileios Karyotis, Symeon Papavassiliou, "Autonomic Network Management and Cross-Layer Optimization in Software Defined Radio Environments", 11(2): 37, *MDPI Future Internet*, 2019.
- **Adamantia Stamou**, "Knowledge Management in Doctoral Education towards Knowledge Economy", *International Journal of Educational Management*, Emerald, 2017.

PUBLICATIONS IN INTERNATIONAL CONFERENCES

- **Adamantia Stamou**, Nikos Dimitriou, Kimon Kontovasilis and Symeon Papavassiliou, "Delay Analysis of Context Aware Management Systems Addressing Multiple Connectivity Opportunities," ADHOC-NOW 2015, Lecture Notes in Computer Science, Springer, vol.9143, pp. 121-133, 2015.

- Jesus Alonso Zarate, Eirini Stavrou, **Adamantia Stamou**, Luis Alonso, Pantelis Angelidis, Christos Verikoukis, "Energy-Efficiency Evaluation of a Medium Access Control Protocol for Cooperative ARQ", International Conference on Communications, Kyoto, Japan, ***IEEE ICC 2011, Best Paper Award,*** 2011.

11. Extended Summary in Greek (Εκτεταμένη Περίληψη στην ελληνική)

Η παρούσα διατριβή πραγματεύεται το ζήτημα της υποστήριξης της συνδεσιμότητας (connectivity management) μεταξύ ετερογενών δικτύων στο πλαίσιο του Διαδικτύου του Μέλλοντος (Future Internet), με βάση τα πρότυπα της αυτονομικότητας (autonomicity) και της επίγνωσης περιβάλλοντος (context-awareness). Στο περιβάλλον του Διαδικτύου του Μέλλοντος, η Πέμπτη γενιά (5G) δικτύων έχει ήδη αρχίσει να καθιερώνεται. Τα δίκτυα 5G αξιοποιούν υψηλότερες συχνότητες παρέχοντας μεγαλύτερο εύρος ζώνης, ενώ υποστηρίζουν εξαιρετικά μεγάλη πυκνότητα σε σταθμούς βάσης και κινητές συσκευές, σχηματίζοντας ένα περιβάλλον ετερογενών δικτύων, το οποίο στοχεύει στο να καλυφθούν οι απαιτήσεις της απόδοσης ως προς την μικρότερη δυνατή συνολική χρονοκαθυστέρηση και κατανάλωση ενέργειας.

Η αποδοτική διαχείριση της συνδεσιμότητας σε ένα τόσο ετερογενές δικτυακό περιβάλλον αποτελεί ανοιχτό πρόβλημα, με σκοπό να υποστηρίζεται η κινητικότητα των χρηστών σε δίκτυα διαφορετικών τεχνολογιών και βαθμίδων, αντιμετωπίζοντας θέματα πολυπλοκότητας και διαλειτουργικότητας, υποστηρίζοντας τις απαιτήσεις των τρεχουσών εφαρμογών και των προτιμήσεων των χρηστών και διαχειρίζοντας ταυτόχρονα πολλαπλές δικτυακές διεπαφές. Η συλλογή, η μοντελοποίηση, η διεξαγωγή συμπερασμάτων και η κατανομή πληροφορίας περιεχομένου σε σχέση με δεδομένα αισθητήρων θα παίξουν κρίσιμο ρόλο σε αυτήν την πρόκληση.

Με βάση τα παραπάνω, κρίνεται σκόπιμη η αξιοποίηση των αρχών της επίγνωσης περιεχομένου και της αυτονομικότητας, καθώς επιτρέπουν στις δικτυακές οντότητες να είναι ενήμερες του εαυτού τους και του περιβάλλοντός τους, καθώς και να αυτοδιαχειρίζονται τις

λειτουργίες τους ώστε να πετυχαίνουν συγκεκριμένους στόχους. Επιπλέον, χρειάζεται ακριβής ποσοτική αξιολόγηση της απόδοσης λύσεων διαχείρισης της συνδεσιμότητας για ετερογενή δίκτυα, οι οποίες παρουσιάζουν διαφορετικές στρατηγικές επίγνωσης περιβάλλοντος, απαιτώντας μια μεθοδολογία που να είναι περιεκτική και γενικά εφαρμόσιμη ώστε να καλύπτει διαφορετικές προσεγγίσεις, καθώς οι υπάρχουσες μεθοδολογίες στην βιβλιογραφία είναι σχετικά περιορισμένες.

Το σύνολο της μελέτης επικεντρώνεται σε δύο θεματικούς άξονες. Στο πρώτο θεματικό μέρος της διατριβής **(Κεφάλαια 2 – 4),** αναλύεται ο ρόλος της επίγνωσης περιβάλλοντος και της αυτονομικότητας, σε σχέση με την διαχείριση της συνδεσιμότητας, αναπτύσσοντας ένα πλαίσιο ταξινόμησης και κατηγοριοποίησης, ώστε να αξιολογηθούν σχετικές λύσεις, με γνώμονα την συνολική βελτιστοποίηση και την αποτελεσματικότητα των αποφάσεων, επεκτείνοντας την τρέχουσα βιβλιογραφία. Στο δεύτερο θεματικό μέρος της διατριβής **(Κεφάλαια 5 – 8),** αναπτύσσεται μεθοδολογία για την ποσοτική αξιολόγηση της απόδοσης λύσεων υποστήριξης της κινητικότητας σε ετερογενή δίκτυα, οι οποίες παρουσιάζουν διαφορετικές στρατηγικές επίγνωσης περιβάλλοντος, ώστε να διαπιστωθεί η καταλληλότητά τους για τα δίκτυα 5ης γενιάς. Ακολουθεί η περιγραφή των επί μέρους κεφαλαίων της διατριβής.

Το **Κεφάλαιο 1**, περιλαμβάνει την συμβολή της διατριβής, καθώς και την δομή. Το **Κεφάλαιο 2**, εισάγει τον αναγνώστη στο θέμα της διαχείρισης της συνδεσιμότητας σε ετερογενή δίκτυα μέσα από το πρίσμα της επίγνωσης εαυτού και περιβάλλοντος και της αυτονομικότητας, ξεκινώντας από τους βασικούς ορισμούς και στη συνέχεια αναπτύσσει το πλαίσιο ταξινόμησης και κατηγοριοποίησης, το οποίο αφορά στις φάσεις της αυτονομικής

διαχείρισης της συνδεσιμότητας, αναλύοντας επιπλέον τις βασικές υπολειτουργίες της κάθε φάσης.

Σύμφωνα με το πρωτόκολλο της IEEE 802.21, η διαδικασία της διαχείρισης της κινητικότητας σε ετερογενή δίκτυα, χωρίζεται σε τρία στάδια. Το πρώτο αποτελεί την έναρξη της μεταπομπής, η οποία περιλαμβάνει την εύρεση δικτύων, την επιλογή δικτύου και την συμφωνία για την μεταπομπή (ανάμεσα στο τρέχων και στο επιλεγμένο δίκτυο). Το δεύτερο περιλαμβάνει την προετοιμασία για την μεταπομπή και πιο συγκεκριμένα την συνδεσιμότητα στο επίπεδο ζεύξης δεδομένων και στο επίπεδο διαδικτύου. Τέλος, το τρίτο, ενέχει την εκτέλεση της μεταπομπής, η οποία περιλαμβάνει την σηματοδοσία, την μεταφορά δεδομένων και την λήψη πακέτων.

Ακολουθώντας μια ελαφρώς εναλλακτική δομή, ώστε να πληρούνται πιο αποτελεσματικά οι ανάγκες της αυτονομικής διαχείρισης μεταπομπών και για να επισημανθούν τα κεντρικά σημεία, η αυτονομική διαχείριση της κινητικότητας μπορεί να χωριστεί σε συλλογή πληροφοριών, η οποία συνδέεται με την βάση γνώσης, ακολουθούμενη από το στάδιο της λήψης απόφασης μεταπομπής, η οποία περιλαμβάνει την έναρξη και την προετοιμασία για μεταπομπή, και τέλος την εκτέλεση της μεταπομπής, η οποία περιλαμβάνει την προετοιμασία για την μεταπομπή και την σηματοδοσία. Σύμφωνα με αυτήν την προσέγγιση, η διαχείριση της κινητικότητας συνάδει με το πλαίσιο της αυτονομικής διαχείρισης και τον κύκλο της αυτονομίας, ο οποίος περιλαμβάνει τα στάδια της παρακολούθησης, της ανάλυσης και του σχεδιασμού, καθώς και της εκτέλεσης.

Πιο αναλυτικά, η διαδικασία της συλλογής πληροφοριών, συλλέγει τις απαραίτητες πληροφορίες οι οποίες είναι απαραίτητες για τις ακόλουθες διαδικασίες της «ανάλυσης» και του «σχεδιασμού». Πιο συγκεκριμένα, η διαδικασία λήψης αποφάσεων έπεται της

διαδικασίας της συλλογής πληροφοριών, η οποία είναι υπεύθυνη για την απόφαση εκκίνησης της μεταπομπής και της επιλογής δικτύου. Άρα θα λέγαμε ότι η αποτελεσματικότητα της απόφασης είναι απόλυτα συνδεδεμένη με την διαθέσιμη πληροφορία, που παρέχεται από την διαδικασία συλλογής πληροφοριών. Πιο αναλυτικά, το περιεχόμενο των πληροφοριών περιλαμβάνει πληροφορία σχετική με τους χρήστες όπως προτιμήσεις, προτεραιότητες, καθώς και το ιστορικό προφίλ του κάθε χρήστη.

Επίσης, το περιεχόμενο των πληροφοριών περιλαμβάνει πληροφορία σχετική με το κινητό τερματικό, όπως η μπαταρία που απομένει, καθώς και παράμετροι σχετικά με την κινητικότητα του ΚΤ, όπως η απόσταση και η θέση του. Ακολούθως, η πληροφορία που συλλέγεται από την διαδικασία συλλογής πληροφοριών αποθηκεύεται στην βάση γνώσης. Η βάση γνώσης κρατά αποθηκευμένο ένα μοντέλο του εσωτερικού και του εξωτερικού περιβάλλοντος, με βάση το κινητό τερματικό και με βάση τα δίκτυα, κάνοντας διαθέσιμο το περιεχόμενο των πληροφοριών στις άλλες οντότητες που διαχειρίζονται το σύστημα και την κινητικότητα. Λαμβάνοντας υπόψιν το κύκλο της αυτονομικότητας και διαχείρισης της κινητικότητας, η βάση γνώσης συνεργάζεται με την διαδικασία της συλλογής πληροφοριών αλλά και την διαδικασία της λήψης αποφάσεων.

Ακολούθως, η διαδικασία λήψης αποφάσεων μπορεί να λεχθεί ότι είναι η βασική διαδικασία που σχετίζεται με την κινητικότητα, καθώς έχει το καθήκον να αναλύει τις πληροφορίες που συλλέχθηκαν από την διαδικασία συλλογής πληροφοριών, καθώς και να σχεδιάζει τις δράσει που θα εξασφαλίσουν την καλύτερη λειτουργία στο σύστημα και στην υποστήριξη της κινητικότητας. Η διαδικασία λήψης αποφάσεων περιλαμβάνει, την έναρξη και την προετοιμασία για μεταπομπή, καθώς και την επιλογή δικτύου, και όλους τους αλγορίθμους που σχετίζονται. Στα αυτονομικά συστήματα, η διαδικασία λήψης αποφάσεων,

περιλαμβάνει επίσης γνωστικές τεχνικές και τεχνικές μάθησης, οι οποίες επιτρέπουν στο σύστημα να ανταπεξέλθει στις ανάγκες του, ενισχύοντας την αυτο-βελτιστοποίηση και την διαδικασία αυτο-ίασης. Πιο αναλυτικά, η διαδικασία λήψης αποφάσεων μπορεί να χωριστεί σε δύο ξεχωριστές διαδικασίες: την επιλογή παραμέτρων, καθώς και την επεξεργασία των παραμέτρων. Πρώτον, η διαδικασία επιλογής παραμέτρων, κάνει χρήση μεθόδων που χρησιμοποιούν την πληροφορία που συλλέχθηκε στο προηγούμενο στάδιο, της συλλογής πληροφοριών, και επιλέγει τις παραμέτρους που απαιτούνται να χρησιμοποιηθούν από τους αλγορίθμους που επιτελούν την επεξεργασία των πληροφοριών, σύμφωνα με κάθε σενάριο. Το σετ των παραμέτρων μπορεί να επιλέγεται από τον χρήστη ή μπορεί να είναι βασισμένο σε κανόνες και πολιτικές, ή και τα δυο.

Τέλος, η διαδικασία της εκτέλεσης, επιτελεί την δράση που σχετίζεται με την εφαρμογή της κάθετης μεταπομπής. Στην παρούσα εργασία επικεντρωνόμαστε στις μεθόδους ελέγχου της μεταπομπής και στις αρχιτεκτονικές διαχείρισης που χαρακτηρίζουν τις λύσεις για αυτονομική υποστήριξη της κινητικότητας, προωθώντας την αυτοδιαχείριση. Η αυτοδιαχείριση αποτελεί μια βασική ιδιότητα της αυτονομικής υποστήριξης της κινητικότητας, καθώς επιτρέπει στο κινητό τερματικό να διαχειρίζεται το ίδιο την λειτουργία του, καθορίζοντας την κατάλληλη στιγμή για μεταπομπή και επιλέγοντας το επιθυμητό δίκτυο. Συμπερασματικά, οι μέθοδοι ελέγχου και οι αρχιτεκτονικές διαχείρισης της κινητικότητας, θα πρέπει να είναι αποκεντρωμένοι (ως ένα βαθμό τουλάχιστον) και να επιτρέπουν στο τερματικό να παίρνει μόνο του τις αποφάσεις που το αφορούν.

Στο **Κεφάλαιο 3**, περιγράφονται τα αυτονομικά χαρακτηριστικά που σχετίζονται με την αυτονομική διαχείριση της κινητικότητας στο πλαίσιο του διαδικτύου του μέλλοντος. Πιο συγκεκριμένα οι αυτονομικές ιδιότητες που σχετίζονται με την διαχείριση της κινητικότητας

είναι η ενημερότητα (awareness) ή αλλιώς επίγνωση εαυτού και περιβάλλοντος, η προσαρμοστικότητα (adaptivity), η ευελιξία (flexibility), η μάθηση (learning) καθώς και η δυνατότητα πρόβλεψης (proactivity). Για τα παραπάνω αυτονομικά χαρακτηριστικά αναλύεται κατά πόσον το καθένα οδηγεί σε βελτιστοποίηση της λειτουργίας του συστήματος. Επιπλέον, θίγεται ένα μείζον θέμα σχετικά με την διαχείριση της συνδεσιμότητας, το οποίο σχετίζεται με την ευρωστία, εννοώντας την λήψη σταθερών και αποτελεσματικών αποφάσεων. Ακολούθως, παρατίθενται για πρώτη φορά στην βιβλιογραφία κανόνες ευρωστίας για λύσεις υποστήριξης της κινητικότητας σε ετερογενή δίκτυα με επίγνωση εαυτού και περιβάλλοντος, καθώς και πώς τα προαναφερθέντα αυτονομικά χαρακτηριστικά συνεισφέρουν στην ευρωστία.

Ακολούθως δίνεται μια περιγραφή των αυτονομικών χαρακτηριστικών ξεκινώντας με την ενημερότητα. Η ενημερότητα συνδέεται με την διαδικασία της παρακολούθησης (monitor), καθώς και με την διαδικασία εύρεσης και επιλογής παραμέτρων προς παρακολούθηση, και αποτελείται από ενημερότητα «εαυτού» και περιβάλλοντος, συλλέγοντας πληροφορία από το τερματικό, τα δίκτυα και τον χρήστη. Η ενημερότητα είναι η βασική ιδιότητα που χαρακτηρίζει ένα αυτονομικό σύστημα και σχετίζεται άμεσα με την βάση γνώσης. Επίσης, το αποτέλεσμα της ενημερότητας είναι η έγκαιρη αντίδραση, η οποία με βάση το πλαίσιο της αυτονομικής διαχείρισης της κινητικότητας, συνδέεται με την «δράση», δηλαδή στην περίπτωσή μας την εφαρμογή της μεταπομπής από ένα δίκτυο σε ένα άλλο.

Στο πλαίσιο της αυτονομικής διαχείρισης δικτύων, η προσαρμοστικότητα συνδέεται με την δυνατότητα του συστήματος να αντιδρά στα τρέχοντα γεγονότα, εξαρτώμενη φυσικά από την ενημερότητα της βάσης γνώσης, και αποφασίζοντας γιατί, πότε, πού και πώς θα επιτελεστεί η ενέργεια της αντίδρασης. Οπότε η προσαρμοστικότητα σχετίζεται με τα στάδια

της «ανάλυσης» και του «σχεδιασμού», από τον κύκλο της αυτονομικότητας. Επιπλέον, η προσαρμοστικότητα μπορεί να συνεισφέρει ως ανάδραση στον κύκλο ελέγχου, η οποία επηρεάζει τις ίδιες τις διαδικασίες της ανάλυσης και του σχεδιασμού, αλλάζοντας για παράδειγμα κάποιον κανόνα ή πολιτική. Για παράδειγμα, η προσαρμοστικότητα θα μπορούσε να επηρεάσει την συχνότητα με την οποία γίνονται οι μετρήσεις στην φάση της συλλογής πληροφοριών, ανταποκρινόμενη στις δυναμικές συνθήκες του περιβάλλοντος. Ακόμη, η προσαρμοστικότητα σχετίζεται με τον δυναμικό χρονοπρογραμματισμό, ελαχιστοποιώντας τον χρόνο που χρειάζεται το σύστημα να πάρει μια απόφαση και κατά συνέπεια βελτιώνοντας την απόδοση του συστήματος. Σύμφωνα με τις τελευταίες επιστημονικές μελέτες, η προσαρμοστικότητα θα μπορούσε να ενισχυθεί με την χρήση αλγορίθμων τεχνητής νοημοσύνης που μιμούνται βιολογικές διαδικασίες, όπως για παράδειγμα η «ευφυΐα σμήνους» (Swarm intelligence), η οποία προσφέρει βελτιωμένη διαχείριση του επεξεργαστικού φορτίου και βελτιώνει και τις τεχνικές δρομολόγησης.

Μια άλλη σημαντική αυτονομική δυνατότητα στο πεδίο της διαχείρισης της κινητικότητας είναι ο βαθμός ευελιξίας του συστήματος. Η ευελιξία σχετίζεται με την ικανότητα του συστήματος να τροποποιεί τις μεθόδους επεξεργασίας, καθώς το σύστημα βρίσκεται σε λειτουργία. Είναι προφανές ότι η δυνατότητα της προσαρμογής της μεθόδου στις μεταβαλλόμενες συνθήκες του περιβάλλοντος, επηρεάζει την πολυπλοκότητα της διαδικασίας λήψης αποφάσεων του συστήματος. Έτσι, ο στόχος είναι να διατηρείται ισορροπία ανάμεσα στην δυνατότητα ευελιξίας του συστήματος και στην επεξεργαστική πολυπλοκότητα. Αν η ευελιξία στηρίζεται σε ένα στατικό σετ παραμέτρων, οι οποίες δίνονται εξ αρχής, τότε πρόκειται για περιορισμένη ευελιξία. Αντιθέτως, όταν το σετ παραμέτρων είναι δυναμικό, δηλαδή μπορούν να προστεθούν νέες παράμετροι ανάλογα με

την περίπτωση, τότε το σύστημα έχει προχωρημένο βαθμό ευελιξίας. Σύμφωνα με την τελευταία μέθοδο, η ευελιξία μπορεί να βασίζεται είτε σε κάποιο πλαίσιο συναρτήσεων (function-based), είτε σε κάποιο πλαίσιο κανόνων (policy-based). Το πλαίσιο των συναρτήσεων περιλαμβάνει τεχνικές όπως η Συνάρτηση Απόφασης (Decision Function) και η Πολυ-Κριτηριακή Απόφαση (Multi-Attribute Decision), και υπολογίζει το αποτέλεσμα της απόφασης με βάση τις παραμέτρους που έχουν επιλεγεί, καθώς και με βάση τα βάρη τους, τα οποία επηρεάζουν ποια παράμετρος έχει μεγαλύτερη επιρροή στην απόφαση. Τα βάρη είναι μεταβλητά, οι παράμετροι όμως όχι, αν δεν υπάρχει ταυτόχρονα και ένα πλαίσιο κανόνων (policy-based). Έτσι, από την άλλη μεριά αν υπάρχει και ευελιξία που στηρίζεται σε ένα πλαίσιο κανόνων, οι κανόνες ενεργοποιούνται με συγκεκριμένες τιμές κατωφλίου, και μπορούν να τροποποιηθούν δυναμικά. Άρα οι λύσεις που υποστηρίζουν ευελιξία που στηρίζεται σε ένα πλαίσιο κανόνων, μπορούν να εξελίξουν περαιτέρω το σύστημα, όμως και πάλι πρέπει να τηρείται η ισορροπία σε σχέση με την πολυπλοκότητα.

Η δυνατότητα μάθησης συνδέεται με την έννοια της γνώσης (cognition). Πιο συγκεκριμένα ένα σύστημα που έχει την δυνατότητα της γνώσης και της μάθησης, μπορεί να θυμάται παρελθοντικές συμπεριφορές (ιστορική μνήμη), όπως π.χ. προβλήματα που προέκυψαν, καθώς και τις λύσεις που ακολουθήθηκαν, επιτρέποντας στο σύστημα να διαθέτει εμπειρία (η οποία συνδέεται με την μνήμη και την βάση γνώσης) και να την χρησιμοποιεί στις αποφάσεις που καλείται να πάρει. Ειδικά σε δυναμικά περιβάλλοντα, η γνώση σχετικά με ακολουθίες συμπεριφορών και ομάδες συνθηκών και σεναρίων μπορεί να βελτιώσει εντυπωσιακά την απόδοση του συστήματος, ανιχνεύοντας ακολουθίες γεγονότων που επαναλαμβάνονται συχνά. Το τελευταίο μπορεί να επιτευχθεί με την χρήση τεχνικών τεχνητής νοημοσύνης, όπως τα νευρωνικά δίκτυα.

Η δυνατότητα πρόβλεψης προωθεί την χρήση προληπτικών μετρήσεων για να διατηρήσει την απόδοση του συστήματος στην παρούσα κατάσταση, προβλέποντας (προσδοκώντας) γεγονότα και την επίδρασή τους στο σύστημα. Η προσδοκία-πρόβλεψη είναι ο ακρογωνιαίος λίθος της προδραστικής υπολογιστικής (proactive computing), προωθώντας ενέργειες στην κατεύθυνση της πρόβλεψης του μέλλοντος. Οι προδραστικές τεχνικές στοχεύουν στην λειτουργία ενημερότητας περιεχομένου (context aware operation), στατιστικής συλλογιστικής (statistical reasoning), καθώς και έξυπνης διαχείρισης δεδομένων (intelligent data-handling). Επιπλέον, συλλέγοντας και αναλύοντας προδραστική πληροφορία, αφορώντας για παράδειγμα την κατάσταση της ζεύξης ή της μπαταρίας, βοηθά το σύστημα να αποφασίσει ποια είναι η κατάλληλη στιγμή για να ξεκινήσει την διαδικασία της μεταπομπής σε άλλο δίκτυο, μειώνοντας τον συνολικό αριθμό μεταπομπών, και μειώνοντας κατά αυτόν τον τρόπο την επιβάρυνση σε σηματοδοσία. Έτσι οι διακοπές της σύνδεσης που μπορεί να βιώσει μια φορητή συσκευή, κατά την διάρκεια κίνησης, μπορεί να αποφευχθεί.

Κατά συνέπεια, τα προδραστικά συστήματα χρησιμοποιούν πληροφορία περιεχομένου ώστε να ανταποκρίνονται πιο γρήγορα και πιο αποτελεσματικά σε συγκεκριμένα ερεθίσματα. Τεχνικές στατιστικής συλλογιστικής όπως τα Hidden Markov Models, οι γενετικοί αλγόριθμοι και οι Bayesian τεχνικές, μπορούν να χρησιμοποιηθούν αντί για τις παραδοσιακές ντετερμινιστικές μεθόδους. Για παράδειγμα, η κινητικότητα ενός χρήστη μπορεί να προβλεφθεί, υπολογίζοντας το ιστορικό στίγμα του χρήστη, και κάνοντας προβλέψεις για το ποια θα είναι η επόμενη τοποθεσία του χρήστη. Επιπλέον, τεχνικές πρόδρασης θα μπορούσαν να χρησιμοποιηθούν για να προ-μετακαλούν (prefetch) δεδομένα, ή να μεταφορτώνουν μαζικά δεδομένα σε έναν κοντινό στο χρήστη εξυπηρετητή (data

staging). Και η τοπική μεταφόρτωση δεδομένων και η μαζική μεταφόρτωση δεδομένων βοηθούν στην υποστήριξη της κινητικότητας του χρήστη. Γενικά, τα προδραστικά χαρακτηριστικά στο πλαίσιο της αυτονομικής διαχείρισης δικτύων, προωθούν την διαθεσιμότητα σε πόρους και δίκτυα, την συμμόρφωση στις συμφωνίες στάθμης (παρεχόμενης) υπηρεσίας (service level agreements), και στην βελτίωσης της συνολικής εμπειρίας για τον χρήστη.

Ο στόχος της αυτο-βελτιστοποίησης είναι να επιτρέπει την ομαλή και αποτελεσματική λειτουργία του συστήματος ακόμη και σε απρόβλεπτα περιβάλλοντα, όπως το διαδίκτυο του μέλλοντος. Η ιδιότητα αυτή προϋποθέτει συνδυασμό χαρακτηριστικών όπως η ενημερότητα, η προσαρμοστικότητα, η ευελιξία, και η προδραστικότητα, οι οποίες οδηγούν το σύστημα σε βελτίωση της απόδοσής του. Η βελτιστοποίηση του συστήματος της διαχείρισης μεταπομπών μπορεί να υλοποιηθεί μέσω της προδραστικής συλλογής πληροφοριών σχετικά με την κατάσταση των πόρων, μετρώντας την παρούσα απόδοση σε σχέση με την ιδανική και εισάγοντας στρατηγικές για μέτρα καλυτέρευσης της κατάστασης.

Ένα μείζον θέμα στην αυτονομική διαχείριση δικτύων, το οποίο αφορά την βελτίωση της απόδοσης, είναι η μείωση και η πρόληψη των μη αναγκαίων μεταπομπών, το οποίο σχετίζεται με την ευρωστία, εννοώντας την λήψη σταθερών και αποτελεσματικών αποφάσεων. Συνεχόμενες οριζόντιες ή κάθετες μεταπομπές μπορούν να οδηγήσουν σε άσκοπη χρήση επεξεργαστικών πόρων και υποβαθμισμένη ποιότητα επικοινωνίας. Ένας μεγάλος αριθμός προσπαθειών κάθετων μεταπομών μπορεί να προκαλέσει αυξημένη καθυστέρηση στην διαδικασία της επεξεργασίας των αιτημάτων για μεταπομπές και μεγάλο ποσοστό απορριπτόμενων κλήσεων. Επίσης, η ενεργειακή αποδοτικότητα επηρεάζεται αρνητικά στην περίπτωση των συχνών μεταπομπών. Συνεπώς, παρατίθεται για πρώτη φορά

στην βιβλιογραφία κριτήρια ευρωστίας για λύσεις υποστήριξης της κινητικότητας σε ετερογενή δίκτυα με επίγνωση εαυτού και περιβάλλοντος.

Πιο συγκεκριμένα, τα κριτήρια ευρωστίας αφορούν την διαχείριση της ποικιλομορφίας και των κανόνων που διέπουν τις αποφάσεις με επίγνωση περιβάλλοντος, περιλαμβάνοντας διαφορετικές τιμές κατωφλίου για κάθε παράμετρο ανάλογα με την κάθε τρέχουσα εφαρμογή, διαφορετική βαρύτητα για κάθε παράμετρο, αντιμετωπίζοντας την ελλιπή πληροφορία και την αβεβαιότητα με π.χ. εισαγωγή ιστορικής πληροφορίας, καθώς και την αντιμετώπιση οριακών περιπτώσεων με την κατάλληλη προσαρμογή τιμών κατωφλίου και υστέρησης. Επίσης, παρουσιάζεται η συνεισφορά των αυτονομικών κριτηρίων στην ενίσχυση της ευρωστίας, προς την κατεύθυνση της συνολικής αυτό-βελτιστοποίησης.

Στο **Κεφάλαιο 4**, αναλύονται σύμφωνα με το προαναφερθέν πλαίσιο ταξινόμησης και κατηγοριοποίησης, έξι χαρακτηριστικές λύσεις για την διαχείριση της κινητικότητας που παρουσιάζουν επίγνωση περιβάλλοντος και αυτονομικές ιδιότητες, μέσα από την τρέχουσα βιβλιογραφία, οι οποίες αξιολογούνται σύμφωνα με τον βαθμό που παρουσιάζουν το κάθε προαναφερθέν αυτονομικό χαρακτηριστικό. Επιπλέον, μελετήθηκε κατά πόσον οι αποφάσεις που λαμβάνονται ως προς την επιλογή του κατάλληλου δικτύου, σύμφωνα με την κάθε λύση, είναι αποτελεσματικές και προτάθηκαν τρόποι βελτιστοποίησης των υπαρχουσών αρχιτεκτονικών, καθώς και προτάσεων προς περαιτέρω ανάπτυξη σχετικών μελλοντικών λύσεων.

Συγκεκριμένα, όπως προέκυψε στα πλαίσια αυτής της μελέτης, ένα σημαντικό στοιχείο που πρέπει να περιλαμβάνουν οι λύσεις διαχείρισης της συνδεσιμότητας, επαφίεται στο γεγονός ότι πολλές δικτυακές διεπαφές είναι ενεργές ταυτόχρονα, οι οποίες εξυπηρετούν διαφορετικές τρέχουσες εφαρμογές. Προς αυτήν την κατεύθυνση, θα είναι καλό να

περιλαμβάνεται στο σετ επιλογής παραμέτρων, πληροφορία σχετικά με τις απαιτήσεις των τρεχουσών εφαρμογών, όπως το εύρος ζώνης, η χρονοκαθυστέρηση, ο αριθμός χαμένων πακέτων και ο λόγος διφυακών σφαλμάτων, ώστε να επιλεχθεί το πιο κατάλληλο δίκτυο για κάθε τρέχουσα εφαρμογή. Ακολούθως, κατά την διάρκεια της διαδικασίας λήψης αποφάσεων, οι απαιτήσεις του χρήστη πρέπει να λαμβάνονται υπόψιν σε μεγάλο βαθμό, όπως επίσης και η ενεργειακή κατανάλωση, το κόστος και η κινητικότητα του τερματικού.

Σχετικά με τις υπό αξιολόγηση λύσεις, για να καλύψουν τις παραπάνω απαιτήσεις, ενδεικτικά αναφέρουμε ότι ορισμένες χρησιμοποιούν Διαδικασία Ιεραρχικής Ανάλυσης, η οποία είναι πολύ αποτελεσματική ως μέθοδος επεξεργασίας παραμέτρων, λόγω της δυνατότητάς της να συνδυάζει και να αξιολογεί πολλαπλά κριτήρια ταυτόχρονα και να ανταποκρίνεται σε πολύπλοκα προβλήματα. Επίσης, η Διαδικασία Ιεραρχικής Ανάλυσης συνδυάζεται με ασαφή λογική, αυξάνοντας την αποτελεσματικότητα του συστήματος, καθώς η ασαφής λογική δύναται να διαχειρίζεται προβλήματα με ασαφή κριτήρια απόφασης. Επιπλέον θα ήταν ωφέλιμο να υπάρχει κάποιο αποθετήριο προφίλ χρηστών με ιστορική πληροφορία, που θα μπορούσε να βελτιστοποιήσει περαιτέρω την ικανοποίηση του χρήστη. Ωστόσο, η χρησιμότητα της Διαδικασίας Ιεραρχικής Ανάλυσης σε πολύπλοκα σενάρια, ίσως δεν είναι το ίδιο επιθυμητή σε απλούστερα σενάρια, όπου μια πιο απλή διαδικασία απόφασης θα μπορούσε να χρησιμοποιηθεί.

Μια εναλλακτική λύση χρησιμοποιεί αυτόματα πεπερασμένων καταστάσεων στην διαδικασία λήψης αποφάσεων, τα οποία επιλύουν όλες τις πιθανές στατικές και δυναμικές συγκρούσεις μεταξύ κανόνων και πολιτικών, και συνεπώς παρέχουν ευστάθεια στο σύστημα και έγκαιρη αντίδραση στα επερχόμενα γεγονότα. Επίσης, λύσεις αυτού του είδους περιέχουν ένα καλά δομημένο, προσαρμοστικό σχήμα για συλλογή πληροφοριών, καθώς και

ευελιξία βασισμένη σε κανόνες στην διαδικασία λήψης αποφάσεων. Ωστόσο, η αυξημένη προσαρμοστικότητα ίσως επιβαρύνει το σύστημα με επιπλέον πολυπλοκότητα. Επομένως, πρέπει να ληφθεί υπόψιν η ισορροπία ανάμεσα στην επιβάρυνση και στο πλεονέκτημα της αυξημένης ευελιξίας. Ειδικότερα, το ζήτημα της αυξημένης πολυπλοκότητας έγκειται περισσότερο σε λύσεις οι οποίες παρουσιάζουν αρκετά κατανεμημένη αρχιτεκτονική, όπου και θα πρέπει να επισημανθεί η ισορροπία που προαναφέρθηκε.

Στο δεύτερο θεματικό μέρος της διατριβής περιλαμβάνει τα **Κεφάλαια 5 - 8**, όπου τίθεται το θέμα της ποσοτικής μέτρησης της συνολικής χρονοκαθυστέρησης των διαφόρων προσεγγίσεων διαχείρισης της συνδεσιμότητας, σε ετερογενή δίκτυα, ως βασικό κριτήριο για την καταλληλότητα σε δίκτυα 5ης γενιάς. Αρχικά, στο **Κεφάλαιο 5**, παρατίθενται τα βασικά χαρακτηριστικά των σχημάτων διαχείρισης κάθετων μεταπομπών από τα διεθνή πρότυπα, καθώς ομάδες διεθνούς προτυποποίησης, όπως το IEEE 802.21 και το 3GPP ANDSF αποτελούν μέρος του θέματος του να παράσχουν ένα ενοποιημένο σχήμα για διαχείριση της ετερογενούς κινητικότητας, προτείνοντας μηχανισμούς που θα μπορούσαν να επιτρέψουν τις κάθετες μεταπομπές.

Σύμφωνα με τα προτυποποιημένα σχήματα, ένας εξυπηρετητής περιεχομένου έχει προταθεί ώστε να αποθηκεύει την πληροφορία περιεχομένου και τις σχετικές πολιτικές, οριζόμενος ως εξυπηρετητής πληροφορίας (ΕΠ). Η πληροφορία που παρέχεται από τον ΕΠ, περιλαμβάνει μια λίστα των δικτύων πρόσβασης που υπάρχουν στην περιοχή που βρίσκεται ο κινητός κόμβος (ΚΚ), πληροφορία τοποθεσίας και πολιτικές σύμφωνες με τον φορέα εκμετάλλευσης, ενώ μπορεί να περιλαμβάνει στατικές παραμέτρους του επιπέδου δικτύου, όπως πληροφορία καναλιού, πολιτικές περιαγωγής μεταξύ διαφορετικών φορέων, κόστη για χρήση του δικτύου κτλ. Και τα δύο πρότυπα εξυπηρετούν μεταπομπές που ελέγχονται από το

τερματικό (mobile controlled), είτε υποβοηθούνται από το δίκτυο (network assisted), είτε υποβοηθούνται από το τερματικό (mobile assisted).

Αναλογιζόμενοι τα κριτήρια των μεταπομπών, η πληροφορία για ποιότητα υπηρεσιών παίζει σημαντικό ρόλο, καθώς τέτοιου είδους πληροφορία μπορεί να κάνει ένα δίκτυο πιο επιθυμητό από κάποιο άλλο, όπως ο φόρτος δικτύου σε σχέση με τις απαιτήσεις του ΚΚ. Πιο συγκεκριμένα, ο παράγοντας του φόρτου αποτελεί μείζων θέμα για την κινητικότητα, καθώς ο ΚΚ μπορεί να μεταπεμφθεί σε ένα άλλο δίκτυο, μόνο στην περίπτωση που το τελευταίο έχει αρκετούς πόρους για τον συγκεκριμένο ΚΚ, απαιτώντας συλλογή δυναμικής πληροφορίας. Οι λύσεις που υπάρχουν στην βιβλιογραφία, περιλαμβάνουν είτε μετα-δραστική συλλογή δυναμικής πληροφορίας, ακολουθώντας τα πρότυπα IEEE 802.21 και 3GPP ANDSF, είτε προ-δραστική, επεκτείνοντας τα πρότυπα. Με βάση τις ήδη υπάρχουσες αρχιτεκτονικές λύσεις, μια ποσοτική αποτίμηση της κάθε λύσης είναι αναγκαία ώστε να αξιολογήσουμε την κάθε προτεινόμενη προσέγγιση, ώστε να συγκριθούν οι δείκτες απόδοσης, όπως η συνολική χρονοκαθυστέρηση από άκρο σε άκρο, η επιβάρυνση λόγω σηματοδοσίας κτλ.

Οι ποσοτικές αποτιμήσεις που υπάρχουν στην βιβλιογραφία μπορούν να χωριστούν σε δύο κατηγορίες. Η πρώτη περιλαμβάνει βασικές προσεγγίσεις με απλή εκτίμηση της χρονοκαθυστέρησης από άκρο σε άκρο, προσθέτοντας τον χρόνο που αναλογεί στις ανταλλαγές μηνυμάτων και βρίσκοντας ποια βήματα στην ακολουθία μηνυμάτων επιβαρύνουν περισσότερο την συνολική χρονοκαθυστέρηση, αλλά χωρίς να λαμβάνουν υπόψιν τους πιθανή συμφόρηση. Η δεύτερη κατηγορία περιλαμβάνει μοντελοποιημένες αποτιμήσεις, περιλαμβάνοντας φαινόμενα συμφόρησης, και πιο συγκεκριμένα, πώς τα διάφορα μηνύματα σχηματίζουν ουρές λόγω της τοπολογίας και της σηματοδοσίας.

Σχετικά με την δεύτερη κατηγορία, οι μοντελοποιημένες αποτιμήσεις δύνανται να τροποποιηθούν ώστε να προσαρμοστούν σε αλλαγές του σχήματος, και με αυτόν τον τρόπο να χρησιμοποιηθούν σε διάφορες αρχιτεκτονικές. Ωστόσο, σχετικά με τις προαναφερθείσες προσεγγίσεις που στοχεύουν στην διαχείριση κάθετων μεταπομπών στην φάση της προετοιμασίας, δεν έχουν ερευνηθεί όλες οι παράμετροι σε επαρκή βαθμό, άρα ένα κοινό μοντέλο συστήματος χρειάζεται ώστε να συμπεριλάβει όλους τους σημαντικούς παράγοντες και να επιτρέψει την λεπτομερή σύγκριση των διαφόρων αρχιτεκτονικών.

Σε αυτό ακριβώς στοχεύουμε με την παρούσα διατριβή, παρουσιάζοντας για πρώτη φορά στην βιβλιογραφία μια γενική αναλυτική μεθοδολογία αξιολόγησης της απόδοσης, στη συνέχεια του **Κεφαλαίου 5**, η οποία μπορεί να προσαρμοστεί σε διαφορετικές τεχνικές συλλογής δυναμικών παραμέτρων, περιλαμβάνοντας και μεταδραστικές και προδραστικές προσεγγίσεις. Πιο συγκεκριμένα το γενικό αναλυτικό μοντέλο (στοχαστικό) που αναπτύχθηκε εστιάζει στην διαδικασία προετοιμασίας των κάθετων μεταπομπών, περιλαμβάνοντας όλους τους παράγοντες που μπορούν να συνεισφέρουν στην συνολική χρονοκαθυστέρηση, όπως ο ρυθμός αιτημάτων μεταπομπής, αριθμός χρηστών, ο αριθμός δικτύων, η πιθανότητα εύρεσης διαθέσιμων πόρων, οι συνθήκες καναλιού, η επεξεργαστική επιβάρυνση, καθώς και η συμφόρηση (μελέτη ουράς). Η συγκεκριμένη μεθοδολογία είναι αρκετά ευέλικτη ώστε να μπορεί να εφαρμοστεί σε διαφορετικές υπάρχουσες αλλά και μελλοντικές λύσεις διαχείρισης της συνδεσιμότητας.

Ακολούθως, στα επόμενα κεφάλαια (**Κεφάλαιο 6** και **Κεφάλαιο 7**) η γενική αναλυτική μεθοδολογία προσαρμόζεται για να μοντελοποιήσει λύσεις με διαφορετική στρατηγική Επίγνωσης Περιβάλλοντος. Συγκεκριμένα, στο **Κεφάλαιο 6**, περιγράφεται και μοντελοποιείται μια μεταδραστική προσέγγιση διαχείρισης της συνδεσιμότητας

ακολουθώντας τα πρότυπα IEEE 802.21 και 3GPP ANDSF. Σύμφωνα με τη συγκεκριμένη προσέγγιση, ο ΚΚ πρέπει να ζητά δυναμική πληροφορία σχετιζόμενη με τους πόρους του κάθε δικτύου, επανειλημμένα, από κάθε δίκτυο ξεχωριστά, μέχρι να βρει δίκτυο με αρκετούς πόρους ώστε να μπορεί να τον φιλοξενήσει.

Λόγου του πιθανά αυξημένου φόρτου σηματοδοσίας που εισάγει η μεταδραστική προσέγγιση, μια εναλλακτική προσέγγιση περιγράφεται και μοντελοποιείται στο **Κεφάλαιο 7**. Το χαρακτηριστικό αυτής της προσέγγισης είναι ότι εισάγει προδραστική συλλογή δυναμικής πληροφορίας, επεκτείνοντας πρότυπα IEEE 802.21 και 3GPP ANDSF, δίνοντας την δυνατότητα στους ΚΚ να παίρνουν την δυναμική πληροφορία απευθείας από μια σχετική οντότητα, η οποία ορίζεται ως αποθετήριο δυναμικής πληροφορίας, χωρίς να χρειάζεται να επικοινωνήσουν με το κάθε δίκτυο ξεχωριστά, κάθε φορά που πυροδοτείται μια μεταπομπή. Συγκεκριμένα, το αποθετήριο δυναμικής πληροφορίας συλλέγει ανά τακτά χρονικά διαστήματα την πληροφορία σχετικά με τον τρέχοντα φόρτο των διαφόρων δικτύων της περιοχής, ώστε σε περίπτωση που υπάρχει ανάγκη για μεταπομπή από ένα ΚΚ, να ενημερώνει όχι μόνο για τα υπάρχοντα δίκτυα της περιοχής αλλά και κατά πόσον είναι διαθέσιμα, σύμφωνα με τις ανάγκες του ΚΚ.

Στο **Κεφάλαιο 8**, παρέχονται τα αποτελέσματα από την μέτρηση της απόδοσης των δύο προσεγγίσεων, ενώ η μεθοδολογία επιβεβαιώνεται από προσομοιώσεις σε κώδικα που αναπτύχθηκε στον NS2 simulator, περιλαμβάνοντας πλήρως επεκτάσιμα μοντέλα προσομοίωσης με μεταβλητό αριθμό τερματικών και δικτύων. Συγκεκριμένα, οι οντότητες έχουν μοντελοποιηθεί ως ουρές, όπου τα πακέτα εισέρχονται και εξέρχονται της ουράς. Η μελέτη περιπτώσεων αφορά δύο σενάρια που άπτονται του τρέχοντος ερευνητικού ενδιαφέροντος και αφορούν ετερογενή δίκτυα, τα οποία παρουσιάζουν υψηλή χωρική

πυκνότητα. Αποδεικνύεται ότι η προδραστική προσέγγιση εμφανίζει καλύτερη απόδοση από την μεταδραστική και στα δύο σενάρια που μελετήθηκαν, πετυχαίνοντας καλύτερη συνάφεια σε σχέση με τις απαιτήσεις των δικτύων 5ης γενιάς. Συγκεκριμένα, η προδραστική προσέγγιση που ακολουθήθηκε εμφανίζει 60%-65% καλύτερη απόδοση ως προς την συνολική χρονοκαθυστέρηση, ενώ χρειάζεται συνολικά 90% λιγότερους επεξεργαστικούς πόρους, παρουσιάζοντας πιθανά οφέλη στην συνολική ενεργειακή κατανάλωση.

Στο **Κεφάλαιο 9**, παρουσιάζονται τα συμπεράσματα από όλη την διατριβή, επιχειρηματολογώντας για τη σπουδαιότητα των εξεταζόμενων ερευνητικών προβλημάτων και των σχεδιαστικών μεθόδων που επιλέχθηκαν για την επίλυση τους, ενώ παράλληλα παραθέτει συγκεντρωμένα τα κύρια συμπεράσματα που ανέκυψαν.

Συνοπτικά τα συμπεράσματα περιλαμβάνουν τα εξής. Καταρχάς, οι αρχιτεκτονικές διαχείρισης της κινητικότητας, θα πρέπει να επιτελούν ενεργή συλλογή πληροφοριών από όλα τα επίπεδα, περιλαμβάνοντας δυναμική πληροφορία που αφορά τις απαιτήσεις των τρεχουσών εφαρμογών, καθώς και να είναι κατανεμημένες, ώστε να επιτρέπουν στο κινητό τερματικό να παίρνει μόνο του τις αποφάσεις, ως ένα βαθμό, δίνοντας την δυνατότητα να διαχειρίζεται αποτελεσματικά το πολύπλοκο περιβάλλον του διαδικτύου του μέλλοντος. Από την άλλη μεριά, οι λύσεις θα πρέπει να μην επιβαρύνουν τα κινητά τερματικά με υψηλή πολυπλοκότητα και κατανάλωση ενέργειας, κάτι που θα μπορούσε να πραγματοποιηθεί με μια υβριδική αρχιτεκτονική προσέγγιση, η οποία επίσης επιτρέπει την συνολική βελτιστοποίηση του συστήματος μέσω της αυτοβελτιστοποίησης των επιμέρους οντοτήτων.

Επιπλέον, η συνάρτηση απόφασης είναι απαραίτητο να συμπεριλαμβάνει ένα ευέλικτο και δυναμικό σετ παραμέτρων, ώστε να μπορεί να υποστηρίζει υπηρεσίες πραγματικού και μη πραγματικού χρόνου. Τέλος, σχετικά με την διαδικασία λήψης αποφάσεων, οι λύσεις

προτείνεται να υποστηρίζουν μηχανισμούς προσαρμογής, όπου διαφορετικοί αλγόριθμοι λήψης αποφάσεων θα μπορούσαν να επιλέγονται με βάση την πολυπλοκότητα του κάθε σεναρίου, ελαχιστοποιώντας την χρήση υπολογιστικών πόρων, την χρήση σηματοδοσίας και χρόνου.

Με την αναλυτική μεθοδολογία που αναπτύχθηκε, αποδείχθηκε ότι η στρατηγική συλλογής πληροφορίας περιεχομένου και ειδικότερα δυναμικής πληροφορίας σχετικά με τους διαθέσιμους δικτυακούς πόρους του κάθε δικτύου παίζει μεγάλο ρόλο στην συνολική χρονοκαθυστέρηση, σχετικά με την διαχείριση των μεταπομπών. Επιπλέον, αποδείχθηκε ότι η τοπολογία (σχετικά με το πόσο πυκνά είναι τα δίκτυα) είναι ένας παράγοντας ο οποίος επίσης επηρεάζει ως ένα βαθμό την συνολική χρονοκαθυστέρηση, καθώς μελετήθηκαν δύο διαφορετικά σενάρια για κάθε προσέγγιση. Συγκεκριμένα η προδραστική προσέγγιση εμφάνισε καλύτερη απόδοση από την μεταδραστική και στα δύο σενάρια που μελετήθηκαν, πετυχαίνοντας μεγαλύτερη συνάφεια με σε σχέση με τις απαιτήσεις των δικτύων 5ης γενιάς. Ταυτόχρονα, αποδείχθηκε ότι η προδραστική προσέγγιση χρειάζεται συνολικά σημαντικά λιγότερους επεξεργαστικούς πόρους, δεδομένου του γεγονότος ότι η οντότητα που συλλέγει προδραστικά την πληροφορία από τα διάφορα δίκτυα, αναλαμβάνει την επεξεργασία των αιτημάτων, αντικαθιστώντας την επεξεργαστική επιβάρυνση από το κάθε ένα δίκτυο, το οποίο συνέβαινε με την μεταδραστική προσέγγιση, παρουσιάζοντας πιθανά οφέλη και στην συνολική ενεργειακή κατανάλωση.

Τέλος προτείνονται ανοιχτά ερευνητικά θέματα για μελλοντική εργασία που είτε θα μπορούσαν να αποτελούν την συνέχεια αυτής της ερευνητικής προσπάθειας, είτε μπορούν να εκμεταλλευτούν την αποκτημένη γνώση προκειμένου να την εφαρμόσουν σε νέους τομείς και δραστηριότητες.

Συγκεκριμένα, τα ανοιχτά ερευνητικά θέματα για μελλοντική εργασία που προκύπτουν από την παρούσα διατριβή, ανάγονται στο γεγονός ότι η αυτονομική διαχείριση δικτύων σε σχέση με την κινητικότητα του χρήστη σε ετερογενή δίκτυα θα μπορούσε να συνδυαστεί με αναπτυσσόμενες τεχνολογίες που συνδέονται με τα δίκτυα 5ης γενιάς, όπως δικτύωση καθοριζόμενη από λογισμικό - Software Defined Networks (SDN), και εικονικοποίηση δικτυακών λειτουργιών - Network Functions Virtualization (NFV), επιτρέποντας την εξέλιξη από τα μοντέλα διαχείρισης δικτύων με βάση την συσκευή, στα μοντέλα διαχείρισης δικτύων με βάση την ποιότητα υπηρεσιών και την επίγνωση περιβάλλοντος (ενημερότητα περιεχομένου), χρησιμοποιώντας την αυτονομική διαχείριση με στόχο την συνολική βελτιστοποίηση των λειτουργιών του δικτύου, και συγκεκριμένα της διαχείρισης της συνδεσιμότητας.

Επιπροσθέτως, τα αυστηρά κριτήρια σε σχέση με την χρονοκαθυστέρηση, απαιτούν πιο αποκεντρωμένες μελλοντικές αρχιτεκτονικές που θα επιτρέπουν επεξεργασία και αποθήκευση δεδομένων στα άκρα του δικτύου, ώστε να αποσοβείται η επικοινωνία με το δίκτυο κορμού. Η αυτονομική διαχείριση δικτύων θα μπορούσε να συνεισφέρει προς την κατεύθυνση της αποκέντρωσης, καθώς εισάγει την αυτό-διαχείριση της λειτουργίας της κάθε σχετικής οντότητας, το οποίο σχετίζεται άμεσα και με την αποκέντρωση της διαχείρισης λύσεων σχετικά με την συνδεσιμότητα. Σχετικά με το τελευταίο, θα μπορούσαν να υποστηριχθούν και σενάρια που επιτρέπουν την επικοινωνία συσκευής με συσκευή, μειώνοντας ακόμα περισσότερο την χρονοκαθυστέρηση και την χρήση δικτυακών πόρων. Η αναλυτική μεθοδολογία που αναπτύχθηκε θα μπορούσε να προσαρμοστεί σε αντίστοιχες μελλοντικές λύσεις, ώστε να συγκρίνονται μεταξύ τους με στόχο την προσαρμογή με τις απαιτήσεις των δικτύων στα πλαίσια του Διαδικτύου του Μέλλοντος.

12. References

[1] A. Galis and A. Gavras, Eds., *The Future Internet - Future Internet Assembly 2013: Validated Results and New Horizons*. Heidelberg: Springer, 2013.

[2] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, "Software defined mobile networks: concept, survey, and research directions," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 126-133, 2015.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What Will 5G Be?," *IEEE Journal On Selected Areas In Communications*, vol. 32, no. 6, pp. 1065-1082, 2014.

[4] A. Kousaridas, C. Polychronopoulos, N. Alonistioti, A. Marikar, J. Mödeker, A. Mihailovic, G. Agapiou, I. Chochliouros, and G. Heliotis, "Future Internet Elements: Cognition and Self-management Design Issues," in *Autonomics '08 Proceedings of the 2nd International Conference on Autonomic Computing and Communication Systems*, 2008, pp. 13-20.

[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions.," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645-1660, 2013.

[6] A. Ahmed, L. M. Boulahia, and D. Gaiti, "Enabling Vertical Handover Decisions in Heterogeneous Wireless Networks: A State-of-the-Art and A Classification," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 776-811, 2014.

[7] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414-454, 2014.

[8] Z. Movahedi, M. Ayari, R. Langar, and G. Pujolle, "A Survey of Autonomic Network Architectures and Evaluation Criteria," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 464-490, 2012.

[9] K. Tsagkaris, M. Logothetis, V. Foteinos, G. Poulios, M. Michaloliakos, and P. Demestichas, "Customizable autonomic network management: integrating autonomic network management and software-defined networking.," *IEEE Vehicular Technology*

Magazine, vol. 10, no. 1, pp. 61-68, 2015.

[10] P. Neves, R. Calé, M. Costa, G. Gaspar, J. Alcaraz-Calero, Q. Wang, J. Nightingale, G. Bernini, G. Carrozzo, A. Valdivieso, L. J. García Villalba, M. Barrose, A. Gravas, J. Santos, R. Maia, and R. Pretog, "Future mode of operations for 5G–The SELFNET approach enabled by SDN/NFV," *Computer Standards & Interfaces*, vol. 54, 2017.

[11] J. A. Wickboldt, W. P. De Jesus, P. H. Isolani, C. B. Both, J. Rochol, and L. Z. Granville, "Software-defined networking: management requirements and challenges," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 278-285, 2015.

[12] "IEEE Standard for Local and Metropolitan Area Networks- Part 21: Media Independent Handover, IEEE Std 802.21-2008, pp.c1-301," Jan. 2009.

[13] "IEEE Standard for Local and metropolitan area networks-Part 21: Media Independent Services Framework," IEEE Std. 802.21-2017, April 2017.

[14] "IEEE Standard for Local and metropolitan area networks-Part 21.1: Media Independent Services," IEEE Std. 802.21.1 -2017, April 2017.

[15] "Access Network Discovery and Selection Function (ANDSF) Management Objects (MO)," 3GPP TS 24.312, V14.1.0 Release 14, June 2017.

[16] S. Barmpounakis, A. Kaloxylos, P. Spapis, and N. Alonistioti, "Context-aware, user-driven, network-controlled RAT selection for 5G networks," *Computer Networks*, vol. 113, pp. 124-147, 2017.

[17] P. Neves, J. Soares, S. Sargento, H. Pires, and F. Fontes, "Context-aware media independent information server for optimized seamless handover procedures," *Computer Networks*, vol. 55, no. 7, pp. 1498-1519, 2011.

[18] B. S. Ghahfarokhi and N. Movahhedinia, "Context-Aware Handover Decision in an Enhanced Media Independent Handover Framework," *Wireless personal communications*, vol. 68, no. 4, pp. 1633-1671, 2013.

[19] Y. Xu, J. Song, J. Li, G.X. Kok, and K. Utsu, "Load-Aware Based Independent Information Services For Heterogeneous Handover," *International Journal of Innovative Computing, Information and Control*, vol. 12, no. 1, pp. 243-262, 2016.

[20] A. Stamou, N. Dimitriou, K. Kontovasilis, and S. Papavassiliou, "Autonomic Handover Management for Heterogeneous Networks in a Future Internet Context: A Survey," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, 2019.

[21] S. Schmid, M. Sifalakis, and D. Hutchison, "Towards Autonomic Networks," *Autonomic Networking, Lecture Notes in Computer Science*, vol. 4195, pp. 1-11, 2006.

[22] B. Jennings, S. Van Der Meer, S. Balasubramaniam, D. Botvich, M. O. Foghlú, W. Donnelly, and J. Strassner, "Towards Autonomic Management of Communications Networks," *IEEE Communications Magazine*, vol. 45, no. 10, pp. 112-121, 2007.

[23] N. Samaan and A. Karmouch, "Towards Autonomic Network Management: an Analysis of Current and Future Research Directions," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 3, pp. 22 - 36, 2009.

[24] K. Tsagkaris, P. Vlacheas, G. Athanasiou, V. Stavroulaki, S. Filin, H. Harada, J. Gebert, and M. Mueck, "Autonomics in wireless network management: Advances in Standards and Further Challenges," *IEEE Network*, vol. 25, no. 6, pp. 41-49, 2011.

[25] P. Makris, D. N. Skoutas, and C. Skianis, "A survey on context-aware mobile and wireless networking: On networking and computing environments' integration," *IEEE communications surveys & tutorials*, vol. 15, no. 1, pp. 362-386, 2013.

[26] M. Z. Asghar, P. Nieminen, S. Hämäläinen, T. Ristaniemi, M. A. Imran, and T. Hämäläinen, "Towards proactive context-aware self-healing for 5G networks," *Computer Networks*, vol. 128, pp. 5-13, 2017.

[27] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Computer Communications*, vol. 31, no. 10, pp. 2607-2620, 2008.

[28] J. Marquez-Barja, C. T. Calafate, J. C. Cano, and P. Manzoni, "An overview of vertical handover techniques: Algorithms, protocols and tools," *Computer Communications*, vol. 34, no. 8, pp. 985-997, 2010.

[29] X. Yan, Y. A. Şekercioğlu, and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks," *Computer Networks*, vol. 54, no. 11, pp. 1848-1863, 2010.

[30] S. Fernandes and A. Karmouch, "Vertical mobility management architectures in wireless networks: A comprehensive survey and future directions," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 1, pp. 45-63, 2012.

[31] M. F. Tuysuz and R. Trestian, "Energy-efficient vertical handover parameters, classification and solutions over wireless heterogeneous networks: a comprehensive survey," *Wireless Personal Communications*, vol. 97, no. 1, pp. 1155-1184, 2017.

[32] P. Bellavista, A. Corradi, and C. Giannelli, "A Unifying Perspective on Context-Aware Evaluation and Management of Heterogeneous Wireless Connectivity," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 337-357, 2011.

[33] M. Louta and P. Bellavista, "Bringing Always Best Connectivity Vision a Step Closer: Challenges and Perspectives," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 158-166, 2013.

[34] A. Stamou, N. Dimitriou, K. Kontovasilis, and S. Papavassiliou, "Delay Analysis of Context Aware Mobility Management Systems Addressing Multiple Connectivity Opportunities," *Ad-hoc, Mobile, and Wireless Networks, ADHOC-NOW 2015, Lecture Notes in Computer Science*, vol. 9143, pp. 121-133, 2015.

[35] K. Mitra, A. Zaslavsky, and C. Åhlund, "Context-aware QoE modelling, measurement, and prediction in mobile computing systems," *IEEE Transactions on Mobile Computing*, vol. 14, no. 5, pp. 920-936, 2015.

[36] Hull R., Neaves P., and Bedford-Roberts J., "Towards situated computing.," in *Proc. 1st Int. Symp. Wearable Comput.*, 1997, pp. 143-153.

[37] R. Fernandez-Rojas and et al, "Contextual Awareness in Human-Advanced-Vehicle Systems: A Survey. ," *IEEE Access*, vol. 7, pp. 33304-33328, 2018.

[38] R.W. Thomas, L.A. DaSilva, and A.B. MacKenzie, "Cognitive networks," in *Proceedings of the First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN*, 2005, pp. 352-360.

[39] J. Strassner, J. N de Souza, D. Raymer, S. Samudrala, S. Davy, and K. Barrett, "The design of a novel context-aware policy model to support machine-based learning and reasoning," *Cluster Computing*, vol. 12, no. 1, pp. 17-43, 2009.

[40] "Autonomic Computing: IBM's Perspective on the State of Information Technology," IBM Research Headquarters (manifesto), 2001.

[41] J. O. Kephart and D. M. Chess, "The Vision of Autonomic Computing," *IEEE Computer*, vol. 36, no. 1, pp. 41-50, 2003.

[42] W. F. Truszkowski, M. G. Hinchey, J. L. Rash, and C. A. Rouff, "Autonomous and Autonomic Systems: A Paradigm for Future Space Exploration Missions," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 3, pp. 279-291, 2004.

[43] J. Strassner, D. Raymer, and , S. Samudrala, "Providing Seamless Mobility Using the FOCALE Autonomic Architecture," *Next Generation Teletraffic and Wired/Wireless Advanced Networking, NEW2AN 2007, Lecture Notes in Computer Science*, vol. 4712, pp. 330-341, 2007.

[44] R. Boutaba and I. Aib, "Policy-based Management: A Historical Perspective," *Journal of Networks System Management*, vol. 15, no. 4, pp. 447-480, 2007.

[45] W. E. Walsh, G. Tesauro, J. O. Kephart, and R. Das, "Utility Functions in Autonomic Systems," in *Proceedings of the IEEE International Conference on Autonomic Computing*, 2004, pp. 70-77.

[46] E. E. Tsiropoulou, A. Kapoukakis, and S. Papavassiliou, "Uplink resource allocation in SC-FDMA wireless networks: A survey and taxonomy.," *Computer Networks*, vol. 1-28, pp. 1-28, 2016.

[47] M. Salehie and L. Tahvildari, "Self-adaptive software: Landscape and research challenges.," *ACM Trans. Autonom. Adapt. Syst.*, vol. 4, no. 2, pp. 1-14, 2009.

[48] E. E. Tsiropoulou, P. Vamvakas, and , S. Papavassiliou, "Joint Customized Price and Power Control for Energy-Efficient Multi-Service Wireless Networks via S-Modular Theory.," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 1, pp. 17-28, 2017.

[49] P. Vamvakas, E. E. Tsiropoulou, and S. Papavassiliou, "A user-centric economic-driven paradigm for rate allocation in non-orthogonal multiple access wireless systems.," *EURASIP Journal on Wireless Communications and Networking*, vol. 129, no. 1, 2018.

[50] E.E. Tsiropoulou, T. Kastrinogiannis, and S. Papavassiliou, "Realization of QoS provisioning in autonomic CDMA networks under common utility-based framework," in *World of Wireless, Mobile and Multimedia Networks & Workshops, WoWMoM 2009*, 2009, pp. 1-7.

[51] E. E. Tsiropoulou, P. Vamvakas, G. K. Katsinis, and S. Papavassiliou, "Combined power and rate allocation in self-optimized multi-service two-tier femtocell networks.," *Computer Communications*, vol. 72, pp. 38-48, 2015.

[52] X. Dong, S. Hariri, L. Xue, H. Chen, M. Zhang, S. Pavuluri, and S. Rao, "Autonomia: an autonomic computing environment," in *IEEE International Conference Proc. in Performance, Computing, and Communications Conference*, 2003, pp. 61-68.

[53] R. Chadha, Y.-H. Cheng, J. Chiang, G. Levin, S. W. Li, and A. Poylisher, "Policy-based mobile ad hoc network management for drama," in *Military Communications Conference - MILCOM 2004*, vol. 3, 2004, pp. 1317-1323.

[54] D. M. Chess, A. Segal, I. Whalley, and S. R. White, "Unity: Experiences with a prototype autonomic computing system," in *ICAC '04: Proc. First International Conference on Autonomic Computing, IEEE Computer Society*, 2004, pp. 140-147.

[55] L. Hua and M. Parashar, "Accord: a programming framework for autonomic applications," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2006, pp. 341-352.

[56] A. Malatras, A. M. Hadjiantonis, and G. Pavlou, "Exploiting context-awareness for the autonomic management of mobile ad hoc networks," *Journal of Network and Systems Management*, vol. 15, no. 1, pp. 29–55, 2007.

[57] A. Bassi, S. Denazis, A. Galis, C. Fahy, M. Serrano, and M. Serrat, "Autonomic internet: a perspective for future internet services based on autonomic principles," in *Proc. 2nd IEEE International Workshop on Modelling Autonomic Communications (MACE)*, 2007.

[58] G. Bouabene, C. Jelger, G. Tschudin, S. Schmid, A. Keller, and M. May, "The autonomic network architecture (ANA)," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 1, 2010.

[59] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: key aspects and survey of handover decision algorithms.,"

IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 64-91, 2014.

[60] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M. S. Alouini, "Velocity-aware handover management in two-tier cellular networks.," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1851-1867, 2017.

[61] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M. S. Alouini, "Handover management in 5G and beyond: A topology aware skipping approach.," *IEEE Access*, vol. 4, pp. 9073-9081, 2016.

[62] H. Ibrahim, H. ElSawy, U. T. Nguyen, and M. S. Alouini, "Mobility-aware modeling and analysis of dense cellular networks with C-plane/ U-plane split architecture.," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4879-4894, 2016.

[63] "Third-Generation Partnership Project, Access to the 3GPP evolved packet core (EPC) via non-3GPP access networks," 3GPP TS 24.302, 2014.

[64] A. Francisco, R. A. Enriquez-Caldera, J. M. Ramirez-Cortes, J. Martinez-Carballido, and E. Buenfil-Alpuche, "Towards a Cognitive Handoff for the Future Internet: A Holistic Vision," in *COGNITIVE 2010 : The Second International Conference on Advanced Cognitive Technologies and Applications*, 2010.

[65] J. M. Kang, J. Strassner, S. S. Seo, and J. W. K. Hong, "Autonomic personalized handover decisions for mobile services in heterogeneous wireless networks," *Computer Networks*, vol. 55, no. 7, pp. 1520-1532, 2011.

[66] P. Vidales, J. Baliosian, J. Serrat, and G. Mapp, "Autonomic System for Mobility Support in 4G Networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 12, pp. 2288-2304, 2005.

[67] Y. Cheng, R. Farha, M. S. Kim, A. Leon-Garcia, and J. W. K. Hong, "A generic architecture for autonomic service and network management," *Computer Communications*, vol. 29, no. 18, pp. 3691-3709, 2006.

[68] M. Rahman, R. Ranjan, R. Buyya, and B. Benatallah, "A taxonomy and survey on autonomic management of applications in grid computing environments," *Concurrency and computation: practice and experience*, vol. 23, no. 16, pp. 1990-2019, 2011.

[69] H. Wang, S. Chen, H. Xu, M. Ai, and Y. Shi, "SoftNet: A Software Defined Decentralized Mobile Network Architecture toward 5G," *IEEE Network*, vol. 29, no. 2, pp. 16-22, 2015.

[70] S. Azodolmolky, P. Wieder, and R. Yahyapour, "Cloud computing networking: challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 54-62, 2013.

[71] A. V. Dastjerdi and R. Buyya, "Fog Computing: Helping the Internet of Things Realize Its Potential," *IEEE Computer*, vol. 49, no. 8, pp. 112-116, 2016.

[72] T. A. Yahiya, G. Pujolle, and A. Beylot, "An Autonomic-Oriented Framework Based IEEE 802.21 for Mobility Management in 4G Networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 15, no. 1, pp. 37-51, 2011.

[73] J. M. Kang, H.T. Ju, and J.W.K. Hong, "Towards autonomic handover decision management in 4G networks," in *IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services*, 2006, pp. 145-157.

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

[75] W. Shen and Q. Zeng, "Cost-Function-Based Network Selection Strategy in Integrated Wireless and Mobile Network," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3778-3788, 2008.

[76] S. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie, "Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 865-881, 2009.

[77] F. Tansu and M. Salamah, "Vertical Handoff Decision Schemes for Heteregeneous Wireless Networks: An Overview," *Recent Trends in Wireless and Mobile Networks*, vol. 84, pp. 338-348, 2010.

[78] K. P. Yoon and C. L. Hwang, *Multiple attribute decision making: an introduction*.: Sage publications, 1995.

[79] Y. Kuo, T. Yang, and G. W. Huang, "The use of grey relational analysis in solving multiple attribute decision-making problems," *Computers & Industrial Engineering*, vol. 55, no. 1, pp. 80-93, 2008.

[80] T. L. Saaty, *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill, 1980.

[81] R. Chai, W. G. Zhou, Q. B. Chen, and L. Tang, "A survey on vertical handoff decision for heterogeneous wireless networks," in *Information, Computing and Telecommunication, YC-ICT'09, IEEE Youth Conference*, 2009, pp. 279-282.

[82] E. Stevens-Navarro and V. Wong, "Comparison between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks," in *IEEE Vehicular Technology Conference*, 2006, pp. 947-951.

[83] E. Stevens-Navarro, Y. Lin, and V. W. Wong, "An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, 2008.

[84] S. A. Sharna, M. R. Amin, and M. Murshed, "An Enhanced-MDP Based Vertical Handoff Algorithm for QoS Support over Heterogeneous Wireless Networks," in *IEEE International Symposium on Network Computing and Applications*, 2011, pp. 289-293.

[85] Y. Wang, P. Zhang, Y. Zhou, J. Yuan, F. Liu, and G. Li, "Handover Management in Enhanced MIH Framework for Heterogeneous Wireless Networks Environment," *Wireless Personal Communications*, vol. 52, no. 3, pp. 615-636, 2008.

[86] L. Ni, Y. Zhu, N. Li, and Q. Deng, "Optimal Mobility-aware Handoff in Mobile Environments," in *IEEE 17th International Conference on Parallel and Distributed Systems*, 2011, pp. 534-540.

[87] J. Baliosian and J. Serrat, "Finite state transducers for policy evaluation and conflict resolution," in *Proceedings of the Fifth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2004)*, 2004, pp. 250-259.

[88] M. Rafiq, S. Kumar, N. Kammar, G. Prasad, G. K. S. Garge, S. V. R. Anand, and M. Hegde, "A Vertical Handoff Decision Scheme for End-to-End QoS in Heterogeneous Networks: An Implementation on a Mobile IP Testbed," in *IEEE National Conference on Communications (NCC)*, 2011, pp. 1-5.

[89] X. Haibo, T. Hui, and Z. Ping, "A novel terminal-controlled handover scheme in heterogeneous wireless networks," *Computers & Electrical Engineering*, vol. 36, no. 2, pp. 269-279, 2010.

[90] K. Vasu, S. Mahapatra, and C. S. Kumar, "QoS Aware Fuzzy Rule Based Vertical Handoff Decision Algorithm for Wireless Heterogeneous Networks," in *IEEE National Conference on Communications (NCC)*, 2011, pp. 1-5.

[91] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. P. Makela, R. Pichna, and J. Vallstron, "Handoff in hybrid mobile data networks," *IEEE Personal Communications Mag.*, vol. 7, no. 2, pp. 34-47, 2000.

[92] N. Nasser, S. Guizani, and E. Al-Masri, "Middleware vertical handoff manager: a neural network-based solution," in *Proceedings of the 2007 IEEE International Conference on Communications (ICC'07)*, 2007, pp. 5671-5676.

[93] I. Lassoued, J. M. Bonnin, and A. Belghith, "Towards an Architecture for Mobility Management and Resource Control," in *Wireless Communications and Networking Conference IEEE Proceedings*, 2008, pp. 2846-2851.

[94] F. Giust, L. Cominardi, and C. J. Bernardos, "Distributed mobility management for future 5G networks: overview and analysis of existing approaches.," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142-149, 2015.

[95] Q. T. Nguyen-Vuong, N. Agoulmine, and Y. Ghamri-Douda, "Terminal-Controlled Mobility Management in Heterogeneous Wireless Networks," *IEEE Communications Magazine*, vol. 45, no. 4, 2007.

[96] Y. Chen, J. Hsia, and Y. Liao, "Advanced seamless vertical handoff architecture for WiMAX and WiFi heterogeneous networks with QoS guarantees," *Computer Communications*, vol. 32, no. 2, pp. 281-293, 2009.

[97] G. Aristomenopoulos, T. Kastrinogiannis, Z. Li, M. Wilson, J. M. González, J. A. Lozano-López, Y. Li, V. Kaldanis, and S. Papavassiliou, "Autonomic Mobility and Resource Management Over an Integrated Wireless Environment – A GANA Oriented Architecture," in *IEEE International Workshop on Management of Emerging Networks and Services*, 2010, pp. pp. 545-550.

[98] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, "Cloud-based augmentation

for mobile devices: motivation, taxonomies, and open challenges," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 337-368, 2014.

[99] A. G. Ganek and T. H. Corbi, "The dawning of the autonomic computing era," *IBM Systems Journal*, vol. 42, no. 1, pp. 5-18, 2003.

[100] R. Want, T. Pering, and D. Tennenhouse, "Comparing autonomic and proactive computing," *IBM SYSTEMS JOURNAL*, vol. 42, no. 1, pp. 129-135, 2003.

[101] J. Puttonen, G. Fekete, J. Makela, T. Hamalainen, and J. Narikka, "Using Link Layer Information for Improving Vertical Handovers," in *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2005, pp. 1747-1752.

[102] T. J. Stewart, "Dealing with Uncertainties in MCDA," in *Multiple criteria decision analysis: state of the art surveys*.: Springer New York, 2005, vol. 78, ch. V, pp. 445-466.

[103] K. Steele, Y. Carmel, J. Cross, and C. Wilcox, "Uses and Misuses of Multi-Criteria Decision Analysis (MCDA) in Environmental Decision-Making," *Risk analysis*, vol. 29, no. 1, pp. 26-33, 2008.

[104] A. Anselin and P. M. Meire, "Multicriteria Techniques in Ecological Evaluation: an Example Using the Analytic Hierarchy Process," *Biological Conservation*, vol. 49, pp. 215–229, 1989.

[105] G. Herath, "Incorporating community objectives in improved wetland management: the use of analytic hierarchy process," *Journal of Environmental Management*, vol. 70, no. 3, pp. 263-273, 2004.

[106] J. M. Kang, J. Strassner, S. S. Seo, and J. Hong, "Autonomic personalized handover decisions for mobile services in heterogeneous wireless networks," *Comput. Netw.*, vol. 55, no. 7, 2011.

[107] M. Kassar, B. Kervella, and G. Pujolle, "Autonomic-oriented Architecture for an Intelligent Handover Management Scheme," in *IEEE Communication Networks and Services Research Conference*, 2008, pp. 139-146.

[108] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," *Proceedings of IEEE*, vol. 121, 1974.

[109] S. Maaloul, M. Afif, and S. Tabbane, "New method of handling weights for a reliability handover decision," in *International Conference on Multimedia Computing and Systems (ICMCS)*, 2014.

[110] R. A. Ribeiro, "Fuzzy multiple attribute decision making: A review and new preference elicitation techniques," *Fuzzy Sets and Systems*, vol. 78, no. 2, pp. 155-181, 1996.

[111] G.A. Di Caro, S. Giordano, M. Kulig, D. Lenzarini, A. Puiatti, and F. Schwitter, "A cross-layering and autonomic approach to optimized seamless handover," in *Proc. of the 3rd Annual Conference on Wireless On Demand Network Systems and Services*, 2006, pp. 104-113.

[112] G. Edwards, A. Kandel, and R. Sankar, "Fuzzy handoff algorithms for wireless communication," *Fuzzy Sets and Systems*, vol. 110, 2000.

[113] N. D. Tripathi, J. H. Reed, and H. VanLandingham, "An adaptive direction biased fuzzy handoff algorithm with unified handoff candidate selection criterion," in *in Proc. 48th IEEE Veh. Technology Conf.*, 1998.

[114] S. Barmpounakis, A. Kaloxylos, P. Spapis, and N. Alonistioti, "COmpAsS: A Context-Aware, User-Oriented Radio Access Technology Selection Mechanism in Heterogeneous Wireless Networks," in *MOBILITY 2014 : The Fourth International Conference on Mobile Services, Resources, and Users*, 2014.

[115] T. Ahmed, K. Kyamakya, M. Ludwig, K. R. Anne, J. Schroeder, S. Galler, and K. Jobmann, "A Context-Aware Vertical Handover Decision Algorithm for Multimode Mobile Terminals and its Performance," in *Proceedings of the IEEE/ACM Euro American Conference on Telematics and Information Systems (EATIS 2006)*, 2006, pp. 19-28.

[116] "Hotspot 2.0 (Release 2) Technical Specification," Wi-Fi Alliance Technical Committee, Hotspot 2.0 Technical Task Group.

[117] "IEEE 802.11u , IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan network specific requirements, Amendment 9: Interworking with External Networks," 2011.

[118] "3GPP, TR 23.865, V12.1.0 , Study on Wireless Local Area Network (WLAN) Network Selection for 3GPP terminals ," Stage 2, Release 12, December 2013.

[119] B. Orlandi and F. Scahill, "Wi-Fi Roaming–Building on ANDSF and Hotspot 2.0," Alcatel-Lucent, Boulogne-Billancourt, France, Tech. Rep. 2012.

[120] 4G Americas, "Integration of Cellular and Wi-Fi Networks ," White Paper 2013.

[121] N. Dimitriou, L. Sarakis, D. Loukatos, G. Kormentzas, and C. Skianis, "Vertical handover (VHO) framework for future collaborative wireless networks," *International Journal of Network Management*, vol. 21, no. 6, pp. 548-564, 2011.

[122] M. Haviv, *Queues: A Course in Queueing Theory*.: Springer, 2013.

[123] P. Billingsley, *Probability and measure*, 3rd ed.: John Wiley & Sons, 1995.

[124] G. Auer, V. Giannini, I. Gódor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J, Gonzalez, C. Desset, and O. Blume, "Cellular energy efficiency evaluation framework," in *IEEE Vehicular Technology Conference (VTC Spring)*, 2011, pp. 1-6.

[125] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers.," in *In Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing.*, 2010, pp. 826-831.

[126] N. Yu, Z. Song, H. Du, H. Huang, and X. Jia, "Dynamic resource provisioning for energy efficient cloud radio access networks.," *IEEE Transactions on Cloud Computing.*, 2017.

[127] C. Han, T. Harrold, S. Armour, and I. Krikidis, "Green radio: radio techniques to enable energy-efficient wireless networks.," *IEEE communications magazine*, vol. 49, no. 6, 2011.

[128] C. Guimaraes et al., "Empowering Software Defined Wireless Networks through Media Independent Handover Management," in *IEEE Global Commun.Conf. (GLOBECOM)*, 2013, pp. 2204–09.

[129] A. Stamou, G. Kakkavas, K. Tsitseklis, V. Karyotis, and S. Papavassiliou, "Autonomic Network Management and Cross-Layer Optimization in Software Defined Radio Environments. ," *Future Internet*, vol. 11, no. 2, pp. 37-55, 2019.

[130] F. Z. Yousaf et al., "Network slicing with flexible mobility and QoS/QoE support for 5G Networks," in *In 2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 1195-1201.

[131] X. Li et al., "A novel optimized vertical handover framework for seamless networking integration in cyber-enabled systems.," *Future Generation Computer Systems*, vol. 79, pp. 417-430, 2018.

[132] A. Gharsallah, F. Zarai, and M. Neji, "SDN/NFV-based handover management approach for ultradense 5G mobile networks.," *International Journal of Communication Systems*, vol. 3831, pp. 1-15, 2018.

[133] V.G. Nguyen, A. Brunstrom, K.J. Grinnemo, and J. Taheri, "SDN/NFV-based mobile packet core network architectures: A survey.," *IEEE Commun. Surv. Tutor.*, vol. 19, pp. 1567–1602, 2017.

[134] I. Chochliouros et al., "Using small cells for enhancing 5G network facilities.," in *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2017, pp. 264-269.

[135] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "FluidRAN: Optimized vRAN/MEC Orchestration.," in *In IEEE INFOCOM 2018-IEEE Conference on Computer Communications* , pp. 2366-2374.

[136] Z. Zhou, S. Mumtaz, K. M. S. Huq, A. Al-Dulaimi, K. Chandra, and J. Rodriquez, "Cloud miracles: Heterogeneous cloud RAN for fair coexistence of LTE-U and Wi-Fi in ultra dense 5G networks.," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 64-71, 2018.

[137] M. Gramaglia, I. Digon, and V. Friderikos, "Flexible connectivity and QoE/QoS management for 5G Networks: The 5G NORMA view.," in *IEEE International Conference on Communications Workshops (ICC)*, 2016, pp. 373-379.