UNIVERSITY OF THE AEGEAN

SCHOOL OF SCIENCES

DEPARTMENT OF INFORMATION & COMMUNICATION SYSTEMS ENGINEERING

PHD THESIS

Context Aware Resource Management for Mobile and Fixed Networking Systems

PRODROMOS K. MAKRIS

KARLOVASSI SAMOS

AUGUST 2013

PHD THESIS

Context Aware Resource Management for Mobile and Fixed Networking Systems

Prodromos K. Makris

Supervision Committee

Committee Chair:

Charalabos Skianis, Associate Professor, University of the Aegean

Committee Members:

Demosthenes Vouyioukas, Assistant Professor, University of the Aegean

Evangelos Pallis, Associate Professor, Technological Educational Institute of Crete

Examination Committee

Charalabos Skianis, Associate Professor, University of the Aegean

Kimon Kontovasilis, Research Director, NCSR Demokritos

Angelos Rouskas, Associate Professor, University of Piraeus

Ioannis Stavrakakis, Professor, National & Kapodistrian University of Athens

Evangelos Pallis, Associate Professor, Technological Educational Institute of Crete

Demosthenes Vouyioukas, Assistant Professor, University of the Aegean

Theodore Zahariadis, Associate Professor, Technical Educational Institute Chalkida

ABSTRACT

"Context", as a research notion, has been invented and roughly exploited in many fields of computer science since 1960s and refers to the general idea that computers can sense, react and adapt their functionalities based on the information they acquire from their environment. In mobile and wireless networking research field, context awareness concepts have been widely adopted as means to provide more intelligent functionalities in terms of context information acquisition, exchange and evaluation, while business logic breakthroughs are being proposed, too. Context-aware resource management is a new research field dealing with ways that traditional resource management algorithms in mobile and wireless networking systems can have more intelligent decision making mechanisms by fully exploiting all context information being available in their geographical environment. Additionally, taking into account the fact that the silos between mobile and fixed networking systems are gradually breaking down, decision making mechanisms have to be realized from an overall system perspective (i.e. for converged mobile and fixed network infrastructures), that is jointly take into consideration mobile and fixed networking systems resources availability for efficient resource management procedures. Finally, during the last years, due to the continuous convergence of computing and networking systems, context awareness concepts appear to be a major research "glue-point" of such kind of heterogeneous environments' integration. Implications of Cloud Computing (CC) paradigm, which are applicable in mobile and wireless networking area are increasingly gaining ground and Mobile Cloud Computing/Networking (MCC/MCN) is an emerging research area introducing itself as the integration of CC into the existing and upcoming 4G HetNet and beyond setups.

In this PhD thesis, novel context-aware resource management schemes and algorithms are proposed for: a) 4G heterogeneous wireless network (HetNet) environments, b) mobile and fixed networking systems' convergence, and c) hybrid/mobile cloud infrastructures, while their performance is evaluated in comparison with related existing state-of-the-art solutions. In a nutshell, this thesis' main contribution is that it introduces several architectural and algorithmic innovations for context aware resource management in convergent mobile and fixed networking systems towards realizing novel building blocks of the next generation mobile networking continuum.

ΠΕΡΙΛΗΨΗ (GREEK)

Ο τίτλος της παρούσας διδακτορικής διατριβής είναι: "Διαχείριση Πόρων με Επίγνωση Πλαισίου για Ενοποιημένα Κινητά και Σταθερά Δικτυακά Συστήματα".

Η έννοια του "γενικότερου/ευρύτερου πλαισίου" (context) στην έρευνα εισήχθη και αξιοποιήθηκε για πρώτη φορά στο χώρο των επιστημών της μηχανικής Η/Υ τη δεκαετία του 1960 και αναφέρεται στη γενική ιδέα ότι οι ηλεκτρονικές συσκευές μπορούν να "αισθάνονται", να αντιδρούν και να προσαρμόζουν τη λειτουργία τους βάσει των πληροφοριών που συλλέγουν από το περιβάλλον τους. Στο πεδίο των κινητών και ασύρματων επικοινωνιών, οι έννοιες που είναι σχετικές με την επίγνωση του πλαισίου (context awareness) έχουν υιοθετηθεί ευρέως για την παροχή εξυπνότερων λειτουργικοτήτων που έχουν σχέση με τη συλλογή, την ανταλλαγή και την αξιολόγηση των πληροφοριών, ενώ έχουν προταθεί και καινοτόμα επιχειρηματικά μοντέλα και λύσεις που βρίσκουν εφαρμογή στην πραγματική αγορά. Η "διαχείριση πόρων με επίγνωση πλαισίου" είναι ένα νέο ερευνητικό πεδίο που ασχολείται με τη βελτίωση των παραδοσιακών αλγορίθμων διαχείρισης πόρων στα κινητά και ασύρματα δικτυακά συστήματα, έτσι ώστε να μπορούν να παίρνουν εξυπνότερες και αποδοτικότερες αποφάσεις λόγω της πλήρους αξιοποίησης των πληροφοριών που βρίσκεται στο περιβάλλον τους (context information). Επιπρόθετα, λαμβάνοντας υπόψιν το γεγονός ότι στις μέρες μας παρατηρείται μία συνεχής ανάγκη για σύγκλιση των κινητών και σταθερών δικτυακών συστημάτων, οι μηχανισμοί λήψης αποφάσεων θα πρέπει να λαμβάνουν υπόψιν τους τη νέα αυτή πραγματικότητα για την αποδοτικότερη διαχείριση δικτυακών πόρων. Τέλος, λόγω του ότι τα τελευταία χρόνια παρατηρείται επίσης μία σύγκλιση μεταξύ των δικτυακών και των υπολογιστικών (computing) συστημάτων, επίγνωση του πλαισίου σημαίνει ότι οι νέοι αλγόριθμοι διαχείρισης πόρων θα πρέπει να λαμβάνουν ταυτόχρονα υπόψιν τους διαθέσιμους δικτυακούς και υπολογιστικούς πόρους για τη μεταφορά οποιασδήποτε υπηρεσίας στις κινητές ηλεκτρονικές συσκευές οποιουδήποτε χρήστη. Έτσι, η εισαγωγή εννοιών του υπολογιστικού νέφους (cloud computing) στα υπάρχοντα ετερογενή ασύρματα δικτυακά περιβάλλοντα 4ης γενιάς αποτελεί μία ιδιαίτερη πρόκληση για την παγκόσμια ερευνητική κοινότητα.

Στην παρούσα διδακτορική διατριβή, προτείνονται καινοτόμες αρχιτεκτονικές και καινοτόμοι αλγόριθμοι για: α) ετερογενή ασύρματα δικτυακά περιβάλλοντα 4ης γενιάς, β) ενοποιημένα κινητά και σταθερά δικτυκά συστήματα, και γ) υβριδικές και κινητές υποδομές υπολογιστικού νέφους (hybrid/mobile cloud infrastructures). Οι προτεινόμενες αρχιτεκτονικές και προτεινόμενοι αλγόριθμοι αξιολογούνται και συγκρίνονται με υπάρχουσες λύσεις από τη διεθνή βιβλιογραφία.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Charalabos Skianis for his mentorship mentality and all his continuous and seamless technical guidance throughout all these years of our academic cooperation. I feel that his life principles and problem solving attitude have broaden my mind horizons and have given me hints for improving my personality. At the end of the day, I feel lucky that I have gained one more real friend in my life.

I would also like to thank University of the Aegean and especially the Department of Information and Communication Systems Engineering for giving me the opportunity to pursue my PhD research and all faculty and administrative members for the overall cooperative and friendly academic environment we have created all these years in the beautiful island of Samos. More specifically, I would like to personally thank all members of Computer and Communications Systems Lab and especially Dimitris Skoutas, Nikos Nomikos and Dimos Vouyioukas for all our everyday technical discussions and fun we had in the office.

Finally, I would like to thank all the members of my family and all my friends, whose continuous support have provided a great social, emotional and fellowship context for me in order to successfully pursue the current PhD thesis.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2
LITERATURE REVIEW AND CONTEXT LIFE CYCLE

CHAPTER 3
CONTEXT AWARE RESOURCE MANAGEMENT IN A 4G HETNET ENVIRONMENT

CHAPTER 4
CONTEXT AWARE RESOURCE MANAGEMENT FOR MOBILE AND FIXED NETWORKING SYSTEMS' CONVERGENCE

CHAPTER 5
CONTEXT AWARE RESOURCE MANAGEMENT IN MOBILE/HYBRID CLOUD INFRASTRUCTURES

CHAPTER 6
CONCLUSIONS & FUTURE WORK

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

4G	Fourth Generation
ABC	Always Best Connected
ATM	Asynchronous Transfer Mode
BES	Best Effort Service
BM	Backhaul Management
BRU	Basic Resource Unit
BS	Base Station
BSM	Backhaul Selection Module
CA	Context Awareness
CA-FEI	Context Aware framework for Femtocells' Efficient Integration of in IP and cellular infrastructures
CABM	Context Aware Backhaul Management
CAMoWiN	Context Aware Mobile and Wireless Networking
CC	Cloud Computing
CE	Capacity Exchange
CIAM	Context Information Acquisition Module
CM	Context Management
CP	Complete Partitioning
CPS	Complete Partitioning Scheme
CS	Complete Sharing
CSC	Common Service Class
CSCI	Community-Sharing Cloud Infrastructures
CSG	Closed Subscriber Group
CSS	Complete Sharing Scheme
DF	Decode and Forward
DoS	Denial of Service
DSAC	Dynamic Service Admission Control
DSL	Digital Subscriber Line
EE	Energy Efficiency
EPC	Evolved Packet Core
EUG	External Users Group
FAS-CP	Fixed Access Scheme with Common Pool
FI	Future Internet
FMC	Fixed Mobile Convergence
FR	Femto-Relay
fUE	Femtocell User Equipment
GoS	Grade of Service
H2H	Human-to-Human
HCC	Hybrid Cloud Computing
HDS	High Demanding Service
HetNet	Heterogeneous Networks
HP	Hybrid Partitioning
HSPA	High Speed Packet Access
IaaS	Infrastructure as a Service
ICI	Inter-Cell Interference
ICT	Information and Communication Technologies
IFW	Integrated Femto-Wifi
IHCI	In-House oriented Cloud Infrastructures
IoT	Internet of Things
IRAC	IaaS Request Admission Control

IRP	Integrated Resource Partition (IRP)
ISAC	Integrated Services Admission Control
ISG	Integrated Service Group
ISP	Internet Service Provider
IUG	Internal Users Group
JCAC	Joint Call Admission Control
KPI	Key Performance Indicator
LDS	Low Demanding Service
LIPA	Local IP Access
LOS	Line Of Sight
LSA	Link Selection Algorithm
M2M	Machine-to-Machine
MCC	Mobile Cloud Computing
MCN	Mobile Cloud Networking
MCRP	Mobile Cloud Resources Provisioning
MNO	Mobile Network Operator
MT	Mobile Terminal
MTC	Machine Type Communications
MTCD	Machine Type Communication Device
MTCR-AC	Machine Type Communication Request Admission Control
mUE	Macrocell User Equipment
OAS	Open Access Scheme
OCI	Outsourcing oriented Cloud Infrastructures
PPA	Periodic Partition Adjustment
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RN	Relay Node
RRM	Radio Resource Management
SINR	Signal to Interference Noise Ratio
SIPTO	Selected IP Traffic Offload
SLA	Service Level Agreement
SPT	Security Privacy Trust
SR	Service Request
UE	User Equipment
USAC	UMTS Service Admission Control
VHO	Vertical Handover
VM	Virtual Machine
VP	Virtual Partitioning
VPB	Virtual Partition for Background services
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WPAN	Wireless Personal Area Network

CHAPTER 1 – INTRODUCTION

In this introductory chapter, a thorough explanation of the thesis' concepts is provided, and the technical scope of the undertaken research is defined. More specifically: a) the general term of context awareness is explained, b) context-aware resource management concepts are analyzed, and c) ways that context awareness rationale points can be utilized for efficient mobile and fixed networking systems' convergence are described. Research motivation and objectives are provided in section 1.4, while main anticipated impact points of the PhD work are given in 1.5. Finally, a thesis roadmap helps the reader to keep track of the work's structural organization.

1.1 The notion of context in ICT systems

"Context", as a research notion, has been introduced and roughly exploited in many fields of informatics and computer science since 1960s and generally refers to the idea that computers can sense, react and possibly adapt their functionalities based on the information they acquire from their environment. The term "context awareness" (CA) was first explicitly introduced in the research area of pervasive/ubiquitous computing in [1] and refers, in general, to the ability of computing systems to acquire and reason about the context information and subsequently adapt the corresponding applications accordingly. During the last decade, increasing interest has been observed on ways to share and exchange context information among remote and heterogeneous CA systems, too. Towards this trend, mobile and wireless systems appear as the most promising and challenging networking research area for the introduction of novel CA functionalities. As a matter of fact, numerous mobile computing devices and plethora of wireless networking technologies have been developed so as the research area of CA computing [2] to be broadened. The ultimate research vision, widely accepted in the international information and communication technology (ICT) community, is that, continuous convergence and fusion of networking and computing systems has to be stressed and "context awareness" notions can become a cornerstone paving the way towards ICT revolution.

The term context is used in everyday life and can be defined as a set of circumstances or facts that surround a particular event, situation or a problem that needs to be solved. ICT research community has been inspired by context awareness situations in real-life occasions and usually novel frameworks, protocols, algorithms and schemes are mapped to abstract concepts related to social life, work, business context, smart city lifestyle etc. For example, regarding everyday social life context, a person has many alternatives in the ways she/he is behaving in her/his everyday social life. For example, her/his behaviour and actions differ when she/he: a) is at home with close family members, b) is at work cooperating with

colleagues, superiors, customers, etc, c) has fun with friends at a café, restaurant, social event, etc, d) enters an unknown building for the first time, and e) visits another country for vacations/work dealing with unfamiliar circumstances related with local cultural/social/economic/political environment. In all these situations, one analyzes the various parameters of her/his social environment, takes into account any prior knowledge and acts accordingly. So, we may conclude that being aware of the everyday social life context, one can prioritize various behaviour manners at different timeframes, while she/he can manage her/his physical/mental/intellectual/financial etc resources more efficiently fulfilling both individual needs and the needs of her/his social environment as a whole.

Another example inspired by everyday work context can complement on the understanding of the notion of context in ICT systems. More specifically, a farmer has a specific plan about each day's tasks (based on his long-term experience), which need to be done in his farm. It is winter and the weather is a crucial factor for all his everyday's activities. In case the farmer is aware of the weather context, he can re-configure his workplan taking into consideration all the work effort going to be undertaken in a timeframe, where weather context information has low-level of uncertainties (e.g. a couple of days or even a week). However, many questions arise when weather context incurs unexpected interactions to other types of context, too (e.g. all employees are not always available, product quality vs. production cost minimization trade-off, etc). Hence, one may conclude that the farmer can make more intelligent use of his available resources and thus drive smarter decision-making procedures, when he is aware of all contextual information affecting the effectiveness and efficiency of his work.

Similarly to the above-mentioned example, one may assume a business context scenario, where an international funding scheme has great experience on successfully developing specific business plans in geographical areas of its interest (e.g. real-estate in touristic areas of special environmental beauty). In this case, being aware of the political, economical, regulatory, social, cultural and environmental context of the investment would be very crucial for the business plan success and its economic sustainability in a long term basis. Conclusively, the entrepreneur team can make more intelligent use of its available fiscal resources and thus drive smarter decision-making procedures, when it is aware of all contextual information affecting the effectiveness and efficiency of its investment.

Going back to a more ICT-oriented paradigm such as the smart city context, nowadays, a huge number of mobile applications are available taking into consideration context information to provide better, seamless and ubiquitous services to end users. For example, a citizen can select and order a taxi according to the driver's and taxi company's reputation/ratings provided by other citizens/friends, who have already used the same taxi service. One can also select the quickest path to go to a place simply by using a context-aware mobile application, which takes into account the current road traffic, the availability of public

transportation vehicles, other unusual events (e.g. strikes, car accident, etc). Smart parking mobile applications are also a latest trend, in which a car driver can easily find a parking slot according to the points she/he has collected and the geographical area she/he wants to park her/his car. Tourism-oriented mobile applications have also been very useful for tourists, who are visiting a city and want to know about available festivals, social events, archaeological sites, restaurants, malls etc. These kind of context-aware applications are using social networking sites to provide reliable information to tourists in order to better satisfy their requests. Conclusively, a smart mobile phone user can use context-aware mobile applications to identify all her/his available options in a specific geographical area and thus make the best choice according to her/his own criteria and interests.

From the real-life scenarios, which were previously described, it is obvious that the notion of context can be exploited in mobile and wireless networking field, too. The abstract strategy of dealing with research problems from an overall system/network perspective can provide many novel solutions in next generation mobile and wireless networks. For example, in a fourth generation (4G) heterogeneous network (HetNet) environment, there is a large amount of context information, which can be exploited by a large variety of network entities (both at the access and core part of the network) in order intelligent decision making procedures to be realized. The typical paradigm of a 4G HetNet environment contains a mobile user being equipped with a handheld mobile device. This mobile user can enjoy a wide variety of mobile services (e.g. voice/video calls, web browsing, mobile gaming, etc) by attaining access to wide variety of heterogeneous wireless access technologies. Each individual mobile user wants to experience the best quality of service (QoS), but this may contradict with the need of the mobile network operators to realize acceptable revenues. Moreover, there is a large variety of key performance indicators (e.g. QoS, system utilization, security, energy, etc), which have to be taken into account in order multi-objective resource management problems to be efficiently tackled.

As a concluding remark, one may assess that there is a need for sedulous research in the field of context aware mobile and wireless networking (CAMoWiN). More specifically, being adequately aware of the context and the context information life cycle, a converged 4G HetNet deployment setting consisting of both mobile and fixed networking subsystems can better deal with emerging research problems. One of these research problems is investigated in the current PhD thesis and refers to context-aware resource management for mobile and fixed networking systems.

1.2 Context Awareness in Resource Management

In this PhD thesis, novel resource management frameworks and algorithms are proposed for converged mobile and fixed networking systems. The novelty of the proposed algorithms is laying on the context awareness notions as those were sententiously described in the previous section.

There had been numerous research works reported in the international literature dealing with resource management problems in computer networks since the early 1980s, when large-scale networking among a large amount of computers was a reality. Both in circuit-switched and packet-switched wireline networks, it had always been of major importance to have efficient partitioning of network resources in order to achieve both QoS guarantees for heterogeneous information flows and high network capacity utilization [3] [4]. At early 1990s, the concept of integrated service packet networks was established and various architectural and algorithmic innovations regarding efficient resource management procedures had been proposed [5]. The basic concept was that given a constrained pool of network resources and heterogeneous information flows (e.g. real-time, best effort, etc), a wireline broadband network system should efficiently manage the available resources according to various bandwidth/capacity partitioning algorithms. Resource management in wide area asynchronous transfer mode (ATM) networks consists the very first good technical approach to dealing with multi-service traffic, capacity partitioning as well as admission control and scheduling algorithms [6] [7] [8] [9].

Once the wireless access technologies came in the foreground (e.g. cellular networks, WLAN, WiMAX, WPANs, etc), the initial approach was to apply state-of-the-art resource management frameworks, as those were standardized in wireline ATM networks. However, from the mobile and wireless networking perspective, uncertainties of the wireless medium pose stiff challenges to the seamless functionality of ubiquitous services and applications and appears in various facets in all layers of the traditional protocol stack [10]. Therefore, cross-layer radio resource management (RRM) approaches, which assume that any kind of information (including context) has to be exchanged between different layers of the traditional protocol stack, have to be adopted so as end-to-end performance can be optimized by adapting each layer against this information [11]. Moreover, given the fact that next generation wireless networks' reality conglomerates many heterogeneous networking systems, cross-system rationale helped in defining ways to exchange uncertain contextual information among heterogeneous entities residing in different systems, following the same rationale points as in cross-layering regarding overall optimal system performance. More specifically, at a certain geographical area, there may be different heterogeneous wireless radio access technologies (RATs), which have different characteristics regarding their

coverage range, the QoS guarantees they offer, the provided level of security, the data rates they can achieve, the pricing model they follow, etc. Given the fact that there is the need for all types of mobile services integration towards offering anything, anytime, anywhere and always best connected (ABC) services to end users, a mentality shift to cross-system and cross-layer approaches was necessary regarding efficient design of RRM algorithms [12].

Because of the fact that modern networks are becoming increasingly dynamic, heterogeneous, less reliable and larger in scale, it is crucial to make them self-behaving, so that minimum human perception and intervention is needed in order to be managed and controlled [13]. Increasing context awareness in autonomic networks is a main cross-cutting challenge for context aware mobile and wireless networking (CAMoWiN) area [14]. More specifically, mobile and wireless networking research community is seeking ways to design a highly optimized system that will support distributed decision making and will incorporate in its architecture reconfigurable and cognitive aspects, in order to make it more autonomous. Similarly, CA computing community is seeking ways to design optimum context models that can uniformly drive autonomic resource management across the spectrum, allowing whole system self-optimization. Besides, as stated in [15], context awareness is a foundation of all self-x properties including self-configuration, self-organization, self-optimization, self-healing etc [16]. Heterogeneous wireless connectivity management is a field where all these self-x functionalities can be applicable [17]. Furthermore, business-aware network management [18] (i.e. correlation of business-level objectives with network resources management), CA security and privacy control [19] [20] (i.e. ways that security and privacy requirements can affect resource management procedures), personalized services [21] (i.e. ways that QoS/QoE provisioning for each individual mobile terminal can affect resource management from an overall system perspective), cloud computing [22] (in terms of efficiently provisioning right services according to the contexts), and other, are some of the latest concepts incorporated in the CAMoWiN paradigm and are directly inter-related with resource management algorithms' design.

In this thesis, we have identified context awareness as a novel feature, which can improve the performance of resource management algorithms for converged mobile and fixed networking systems. That is, a resource management algorithm can exploit all available context information in order to make better decisions regarding a constrained pool of network resources. However, given the fact that context information can be in various forms (e.g. raw data, low-level, high-level context, etc) and be available at remote and multiple network entities at given time instances, a resource management algorithm should be aware of the whole context information life cycle in order to be able to exploit it in an effective and efficient manner. Conclusively, CAMoWiN research field expands cross-system and cross-layer optimization approaches as 4G heterogeneous networks are continuously increasing

their market share in mobile and wireless communications. According to the real-life scenarios described in the previous section, table 1.1 provides a mapping of situations that show ways that context awareness concepts can assist real-life decision-making and network resources management procedures.

Table 1.1: **Mapping of real-life decision-making and network resources management procedures regarding context information exploitation**

Context awareness in real-life occasions	Context awareness in resource management algorithms
Home with close family members	High QoS for high-priority services
Boss/chief/head of department requests	High-priority user groups
Cooperating with colleagues to achieve team objectives	Consider both each user QoS and system capacity utilization
Enter a (crowded) place for the first time and deal with unfamiliar circumstances	Compromise with minimum QoS, when context is bad/unknown (e.g. cell-edge users, non CSG femtocell users, etc)
Farmer's decisions according to weather predictions	Prediction-based capacity partitioning techniques
One individual group's actions affecting the interests of other people	Protect a whole HetNet system from interference-related QoS degradation
Crisis management situations	Exploit special contextual information to provide acceptable QoS metrics in overload-state situations
Project manager deals with budget and human resources over/under-estimations	Extend/shorten available pool of resources by dynamically changing the network topology

Resource management frameworks in beyond 3G and 4G networks are focusing on the radio access part because this is the prevailing bottleneck in comparison with the backhaul resources, whose pool were assumed to be noticeably larger. However, during the last years a densification of cells is observed (i.e. femtocells, picocells, relays, body area networks, etc) and thus small cell traffic is experiencing a continuous increase rendering network backhaul resources management a crucial research problem, too. For example, this happens in cases where femtocells are attached to xDSL connections (i.e. the backhaul capacity equals some tenths of Mbps bandwidth or even less), while 4G radio access bandwidth can reach even higher values in specific HetNet deployment settings. Based on the above challenges, novel resource management algorithms should be designed in order both radio and backhaul capacity bottlenecks to be taken into consideration. This problem becomes even more complex as 4G HetNet architectures can dynamically change, while the available pool of resources can also considerably change in short time intervals (e.g. wired or wireless backhaul) providing various thresholds of QoS to differentiated user groups and mobile services.

Additionally, cloud computing is expected to revolutionize state-of-the-art mobile and wireless networking technologies providing means of realizing the vision of complete

networking and computing environments' integration [23]. As a result, efficient mobile cloud computing (MCC) resource management frameworks should simultaneously take into consideration both: a) wireless/radio access resources pool aiming at always-best connectivity contexts and b) computing resources pool for data processing/storage aiming at flexible virtualized infrastructure sharing solutions. More specifically, the novelty in this approach lays in the fact that a context aware resource management algorithm should jointly manage radio and computing resources rather than confronting the problem as two independent resource management sub-problems [24].

In this PhD thesis, context aware resource management algorithms are designed and developed for innovative network deployment settings, while their performance is evaluated in comparison with state-of-the-art algorithms and capacity partitioning techniques from the international literature.

1.3 Context Awareness for Mobile and Fixed Networking Systems' Convergence

Fixed Mobile Convergence (FMC) came in the research foreground at the early 2000s as an initial approach to realize 4G networks [25] [26]. FMC concepts provided a transition point in the telecommunications industry that aim to ultimately fuse the distinctions between fixed and mobile networks, providing a superior experience to end users by creating seamless services using a combination of fixed broadband and local access wireless technologies to meet their needs in homes, offices, other buildings and on the go. In general, as long as there is a continuously increasing amount of mobile data traffic, this has (by all means) to efficiently traverse across existing fixed backbone wired networks of mobile operators. As operators had their own backhaul network in order to manage their own mobile traffic generated in the radio access part of their mobile infrastructure, all network capacity problems were focused on the wireless access part and not in the operators' packet core network. As the demand for mobile data traffic was exponentially growing and new heterogeneous wireless access technologies were claiming their market in the telecommunications industry, a more liberalized mentality shift has taken place regarding mobile data traffic routing throughout the whole mobile and fixed networking infrastructures. More specifically, the mobile data traffic generated in a specific geographical area can be handled in various ways according to the environmental and architectural context of the 4G Hetnet topology at a given time interval. For example, femtocells is a new emerging wireless technology, where femtocell base stations (BS) are directly connected to a xDSL backhaul link. So, the case of femtocells interworking with pre-existing wireless networks (e.g. WLAN, WPAN, etc) and wireline networks (e.g. ethernet) poses more challenges due to the sharing of the same backhaul capacity. While the femtocell inherits the QoS mechanisms of cellular networks and is able to provide a reliable resource

management situation, this does not apply to the IP-based networks and that may drastically affect the performance of the femtocell. The problem can become more complex when the femtocell owner wants to take advantage of the entire xDSL backhaul capacity for which has paid, while other mobile terminals in the vicinity of the femtocell are forced to communicate with a macrocell BS, incurring thus interference management issues, which are often directly interrelated with overall system capacity deterioration problems.

Nowadays, there is a trend in having integrated service routers (e.g. small cell gateways, machine-to-machine gateways, etc), which consist an aggregation point for mobile data traffic. As a result, integrated QoS provisioning and efficient resource management solutions are needed in order to achieve an efficient integration of small cells in IP and cellular infrastructures [27]. The idea of combining WiFi, femtocell, router and DSL modem in a single box was first introduced by NETGEAR in 2008 and since then many vendors and operators have encouraged their fixed line and mobile teams to combine their strategies, breaking down the silos between them. With everything integrated into a single gateway, service providers can also better understand the small cell network topology, the corresponding device status and configuration helping them to look beyond small cell technology as a simple remedy for poor coverage in indoor environments. Conclusively, small cell gateway can be considered as an ideal network entity that can take into consideration both radio and backhaul resource requirements and possible bottlenecks and thus appropriately coordinate the overall resource management procedure for converged mobile and fixed networking systems. Moreover, machine-to-machine (M2M) gateways are also gaining their own market share, as according to conservative estimations, cellular M2M connections are expected to overcome 200 million in 2013, while trillions of Machine Type Communication Devices (MTCDs) will exist by the end of the year 2020 [28]. As a result, combined human-to-human (H2H) and M2M traffic needs to be efficiently handled and routed throughout converged mobile and fixed networking infrastructures using context awareness rationale [29].

Due to increasingly higher data rates achieved via the wireless access network technologies (i.e. 4G LTE and LTE-A), wireless backhaul link capacities may often be larger in future HetNet scenarios. That is, in HetNets, it will be required to have flexible network topologies, which will utilize both wired and wireless backhaul alternatives to route the large amounts of mobile data traffic. Hence, in situations where a wired backhaul becomes a bottleneck in a specific local geographical area, other wireless backhaul solutions should be activated. Furthermore, in cases where multiple backhaul alternatives are available, efficient decision making procedures should take place according to the current 4G HetNet context.

In this PhD thesis, a novel approach of having femtocells operating as relays (i.e. femto-relay concept) is investigated and the gains that femto-relays can introduce to a 4G and beyond

environment due to their two-fold functionality as both femtocells and relays are presented. The main differentiation of femto-relays is that their backhaul consists of a combination of wired and wireless communication links, which consider routing of mobile data traffic either through xDSL connection and the Internet (i.e. wired backhaul) or directly to the donor BS (i.e. wireless backhaul) that offers the best signal to interference noise ratio (SINR). Consequently, context awareness concepts in this kind of 4G HetNet deployment settings can also assist mobile and fixed networking systems' convergence.

FMC is a research notion, which has been mainly utilized from networking perspective during the last years. However, nowadays, because of the need for integrated networking and computing environments, flexible virtualized infrastructure sharing solutions have to be combined with traditional RRM solutions for mobile/wireless networks. Hence, RM modules residing in hybrid cloud infrastructures should cooperate with RM modules residing at the wireless/access part of the network. Context awareness can play a critical role in this kind of architectural innovations. A main contribution of this PhD thesis is that in the emerging integrated networking and computing continuum, state-of-the-art resource management frameworks and techniques have to be enhanced in order to confront the related research challenges from both networking and computing perspectives simultaneously. For example, for a given context-aware mobile application, there may be enough networking resources to achieve the required QoS constraints but there may also be computing resources outage, which can lead to application's failure. Consequently, awareness of the overall context in a FMC environment can considerably enhance key system performance indicators.

In this PhD thesis, the need for innovative resource management frameworks for mobile and fixed networking systems' convergence is outlined. FMC concepts are extended towards the realization of integrated networking and computing environments, where overall context awareness can decisively assist in optimal and joint management of both networking and computing resources.

1.4 Motivation and Objectives – Research Scope

In this section, the research motivation and the associated technical objectives of the PhD thesis are described. Prior to that, in figure 1.1, a high-level representation of PhD thesis scope of research is described. As shown in the figure, there are three main architectural pillars namely: a) 4G HetNet environment, b) mobile and fixed networking systems' convergence, and c) hybrid cloud infrastructure.

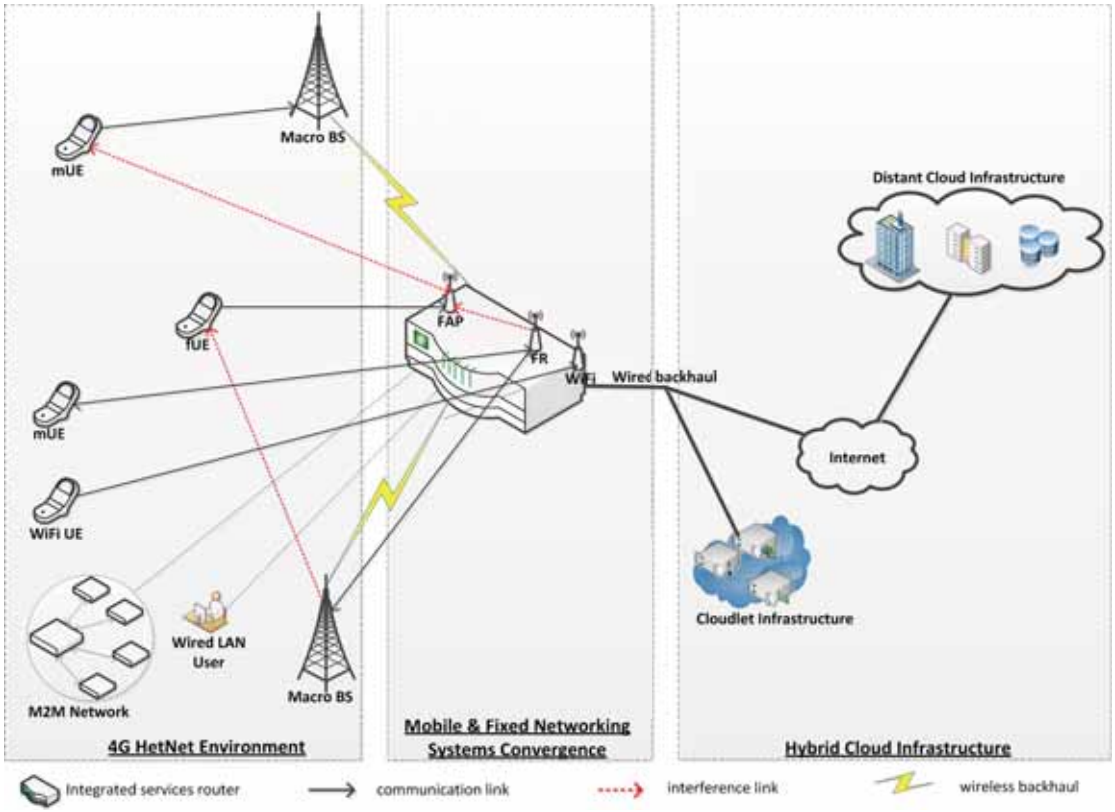

Figure 1.1: A high-level depiction of the PhD thesis scope of research

Regarding the 4G HetNet environment, HetNet concepts mainly deal with problems such as dynamic leveraging of the existing cellular network topology and increasing the proximity between the access network and the end users [30]. This can be done by deploying multi-tier cellular infrastructures, that is, in order to reduce the distance between transmitter and receiver, a HetNet combines macrocells with low-power nodes, such as microcells, picocells, femtocells and relays. Speaking in technical terms, all the above considerations can be realized only if HetNet environments follow certain basic context-aware mobile and wireless networking (CAMoWiN) principles such as those discussed in [23]. More specifically, the more accurate and up-to-date context information is acquired, exchanged and evaluated by all HetNet entities, the better decision making procedures will take place regarding the connectivity options of mobile terminals (MTs) from an overall HetNet environment perspective. Femtocells, while they are able to enhance the coverage and capacity of a cellular network with minimum cost by utilizing existing IP infrastructures, their unstructured placement may also cause major interference-related problems [31]. However, if the femtocell operation is assisted by relays (i.e. femto-relays denoted as FRs), then the signal transmissions to/from the macrocell can be moved closer to the user equipment (UE). As a result, transmitting with better channel conditions leads to transmission power reduction in both uplink and downlink further reducing inter-cell interference (ICI) between the macrocell and the femtocells. Consequently, femtocell technology, if fully exploited, can lead to

improvements in interference reduction and data rate means, not only for the femtocell user equipment (fUEs) but also for nearby macrocell user equipments (mUEs). In addition to cellular users, other types of UEs may reside in a 4G HetNet setup as heterogeneous wireless and wireline network technologies may coexist in the same geographical area [32]. Therefore, MTs have many wireless access alternatives and thus can be connected to the best network technology with the prerequisite that overall system's key performance indicators (KPIs) are not violated. Finally, machine type communications (MTC) is an emerging technology, which is already playing a critical role in Future Internet evolution and is expected to be a key player in 4G and beyond HetNet environments.

As of the second architectural pillar depicted in figure 1.1 (i.e. mobile and fixed networking systems' convergence), we assume a mobile data traffic aggregation point, where heterogeneous information flows have to be served according to their individual QoS characteristics and the available network's backhaul resources constraints. This kind of devices is expected to prevail in the market and nowadays can be found in the international literature and real market brochures as integrated services routers, small cell gateways, M2M/MTC gateways, etc. Their main operational role is to assist mobile and fixed networking systems' convergence by providing efficient resource management procedures thus allowing heterogeneous mobile data traffic streams to be seamlessly traversed throughout the entire network infrastructure. The main technical challenge is to make the heterogeneity of access technologies transparent to the end user. More specifically, compared to the general case where the communicating networks are independent, the case of femtocells interworking with pre-existing wireless/wireline networks poses more challenges due to the sharing of the same backhaul capacity. Therefore, while a user is practically able to initiate the same service through multiple network interfaces, he is allocated capacity from the same capacity pool and thus integrated QoS provisioning approaches are required [33]. The resource management modules, which reside at the integrated services router are able to communicate and cooperate with other resource management modules residing in hybrid cloud computing (CC) infrastructures (see right-hand side of figure 1.1) as well as with all MTs, UEs, MTC devices (MTCDs), cellular BS and wireless RATs' access points residing at the 4G radio HetNet environment depicted on the left-hand side of figure 1.1. As a result, context awareness concepts can be adopted for integrated services router modules and thus more intelligent decision making procedures can take place. In other words, the aforementioned modules are able to employ CA signaling protocols in order to communicate with network modules residing at both MTs and core network's sides, and thus fine-tune the resource management procedures from an overall system's perspective.

Regarding the last architectural pillar (i.e. hybrid CC infrastructures), CC infrastructures are typically based on virtualized environments to allow physical infrastructure to be shared by

multiple and diverse end users. However, the efficient sharing of a cloud infrastructure can be performed through user/service-centered resource management procedures, which should also be flexible enough to adapt to the various real market cloud deployment scenarios. Several conflicting parameters such as the type of the hybrid cloud infrastructure being deployed, multiple user priority groups, security, energy efficiency and financial costs should also be taken into account. Thus, aiming to deal with all these emerging resource management trade-off problems, highly customizable infrastructure sharing approaches for hybrid CC environments are proposed. For example, summarizing the basic capacity management objectives that a virtualized cloud environment should satisfy, these can be: a) load balancing, b) adaptive resource allocation, c) flexibility, d) security, e) fine-grained resource control, f) continuous resources availability, g) high system utilization performance, and h) deal with unreliably and unpredictably excessive computing resources [34]. In addition to computing resources' QoS provisioning, joint management of networking resources should also be employed. For example, we study the future MCC continuum as an optimal combination of: a) distant CC infrastructures (e.g. public/community clouds), b) proximate cloudlet infrastructures (e.g. private clouds), and c) communicating objects and smart mobile devices. In this kind of deployment settings and due to the explosion of mobile applications market, the average mobile user demands for computing/storage power is much higher than the one that can be supported by an average MT and this gap is continuously growing. As a result, mobile users need to access servers located in virtualized CC infrastructures in order to meet their increasing functionality demands. A cloudlet infrastructure consists of a cluster of servers well-connected to the Internet and available for use by nearby MTs. A cloudlet can be contained in a MCC hotspot together with a wireless access point comprising thus a datacenter-in-a-box concept. These MCC hotspots can be placed in cafes, shopping malls, airports, stadiums, museums, city squares, campuses, train stations, etc. In these environments, mobile users may meet the demand for real-time interactive responses for specific-purpose context-aware mobile applications by low-latency, one-hop and high-bandwidth wireless access to the cloudlet. The problem with cloudlets is their restricted computing/storage capabilities, taking into account that they have to (by priority) meet the QoS demands of numerous users for specific-purpose context-aware mobile applications in a restricted geographical space. Consequently, for general e-* related MCC services (i.e. e-government, e-health, e-banking, e-commerce, e-learning, etc), the access to a community cloud infrastructure would be more appropriate for mobile users requesting corresponding specific e-* mobile applications. A community cloud or micro-cloud (cf. microcells in mobile networking continuum) is controlled and used by a group of industrial/institutional organizations, which have common or shared financial, security and legal objectives. Despite of their high availability and targeted applicability to specific mobile users' demands, high

WAN latency is an inevitable trade-off. The same problem is valid for public clouds. The main difference is that Amazon/Apple/Google/Microsoft etc large data centers can provide virtually infinite computing/storage capabilities for any type of MCC service.

1.4.1 Research Motivation

Having already defined the technical scope of this PhD thesis research, there are three main motivation drivers, which have directed the research efforts, summarized as follows:

1) To handle research problems in the mobile and wireless networking field from an overall network system's perspective (cf. the notion of context in ICT systems sententiously described in 1.1).

2) To enhance state-of-the-art resource management algorithms in order to be applicable to 4G heterogeneous network architectures (cf. context awareness in resource management algorithms described in 1.2).

3) To investigate context awareness aspects as a major research "glue-point" of networking and computing systems' integration (cf. context awareness for mobile and fixed networking systems' convergence described in 1.3).

1.4.2 Research Objectives

Following research motivation drivers and a thorough study of related state-of-the-art issues, this PhD thesis introduces several architectural and algorithmic innovations for context aware resource management in mobile and fixed networking systems. More specifically, the research objectives can be summarized as follows:

1) To define the novel research field of "Context-Aware Mobile and Wireless Networking" (CAMoWiN), identify all CA functionalities of a typical CAMoWiN system and the phases of context information life cycle.

2) To propose a context-aware resource management framework for coexisting small cell, wireless and wireline networking environments and evaluate its performance in comparison with state-of-the-art capacity partitioning schemes.

3) To provide an integrated QoS provisioning framework for small cell/MTC gateways.

4) To study the novel "femtocell as a relay" concept for 4G HetNet deployments and propose context-aware resource management algorithms to deal with: a) various types of interference mainly caused by unplanned small cell deployments, and b) limited and QoS-unreliable wired backhaul links.

5) To propose a context-aware framework for the efficient integration of small cells in IP and cellular infrastructures and evaluate the overall performance of a 4G HetNet environment in terms of QoS provisioning, energy saving and data rate enhancements.

6) To propose a context-aware mobile cloud resources management approach to jointly handle: a) wireless/radio access resources pool aiming at always-best connectivity contexts, and b) computing resources pool for data processing/storage aiming at flexible virtualized infrastructure sharing solutions.

7) To propose a context-aware resource management scheme for hybrid cloud computing environments and evaluate its performance in terms of QoS differentiation for multiple user groups and application-level security QoS provisioning.

1.5 Anticipated Impact

The impact that the aforementioned novel contributions of this PhD thesis may have in the international community has been a serious concern of the undertaken research throughout the lifetime of the PhD work. In order to ensure high impact factors for the PhD work in the short and long term future, we have followed some of the widely recognized good practices such as:

a) Publish PhD research results in good quality and high-impact factor international journals and conferences.

b) Exploit active participation in large-scale (EU-level) research projects dealing with up-to-date research challenges in the "future networks" field, communicate results in high quality audiences/partners ranging from academic universities and research institutes to small/medium enterprises (SMEs) and large companies in the telecommunications sector.

c) Find good balancing points between academia and industry-oriented impact of the research results.

d) Work on real-market implementation problems and apply context awareness concepts in ongoing projects and/or project proposals requesting funding for further enhancing the quality of research outcomes in the future.

e) Participate and follow-up latest research trends (e.g. EU research agenda) and real market products, which may adopt architectural and/or algorithmic solutions proposed in this PhD context.

Working on the above-mentioned dissemination and exploitation strategy plan and refining it throughout the whole PhD project's lifetime, we can summarize some main anticipated impact points of our work as follows:

1) Provide novel research insights about networking and computing systems' integration. For example, we have dealt with an emerging research field called mobile cloud computing/networking, which is expected to prevail at the field of mobile and wireless networking in the upcoming years.

2) Provide novel research insights for jointly tackling resource management problems in a fixed-mobile networking systems convergence context. More specifically, we have enhanced state-of-the-art RM frameworks including radio, fixed networking and computing resources management in the same multi-objective decision making procedures.

3) Provide prototype implementation for RM modules applicable in integrated services routers, small cell and M2M gateways, which are expected to experience a wide market adoption in the upcoming years.

4) Propose novel 4G HetNet deployment scenarios being able to boost beyond 4G market opportunities and business models (e.g. femto-relay concept). For example, mobile service providers can increase their revenues by exploiting promising femto-relaying capacity gains towards delivering novel and even more resource-consuming context-aware mobile applications. Vendors/manufacturers can benefit from developing elegant H/W solutions and customer-friendly small cell products. Research and retail-oriented SMEs can also benefit from the envisioned femto-relay market growth, while citizens will be provided with the wider selection of mobile applications with better QoS, seamless service provisioning and service outages minimization.

5) Contributions to the undergoing 4G networks and beyond research in terms of overall network system's capacity increase, spectral efficiency, energy efficiency, etc following reliable performance enhancement and expected impact targets (e.g. indicated by Digital Agenda for ICT research in Europe, EU framework program 7 and 8 for periods 2007-2013 and 2014-2020 respectively).

1.6 Publications List

Part of the contribution of this thesis has been published/accepted for publication/is under review in widely recognized and high-impact factor journals and has been presented in IEEE international conferences of the telecommunications' field.

1.6.1 Journals/Magazines

- 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(1), pp. 362-386, 2013.
- P. Makris, N. Nomikos, D. N. Skoutas, D. Vouyioukas, C. Skianis, J. Zhang and C. Verikoukis, "A Context Aware Framework for the Efficient Integration of Femtocells in IP and Cellular Infrastructures", Springer EURASIP Journal on Wireless Communications and Networking, Special Issue on Small Cell Cooperative Communications, 2013:62, http://jwcn.eurasipjournals.com/content/2013/1/62.

- D. N. Skoutas, P. Makris and C. Skianis, “Optimized Admission Control Scheme for Coexisting Femtocell, Wireless and Wireline Networks”, Springer Telecommunication Systems Journal, available online, 2013, http://link.springer.com/article/10.1007%2Fs11235-013-9703-4
- A. Bourdena, P. Makris, D. N. Skoutas, C. Skianis, G. Kormentzas, E. Pallis and G. Mastorakis, “Joint Radio Resource Management in Cognitive Networks: TV White Spaces Exploitation Paradigm”, IGI Global for Evolution of Cognitive Networks and Self-Adaptive Communication Systems, June 2013.
- N. Nomikos, P. Makris, D. N. Skoutas, D. Vouyioukas and C. Skianis, “Wireless Femto-Relays: A New Model for Small Cell Deployments”, submitted in IEEE Wireless Communications Magazine, July 2013.

1.6.2 Conferences/Workshops

- P. Makris, D. N. Skoutas, P. Rizomiliotis and C. Skianis, “A User-Oriented, Customizable Infrastructure Sharing Approach for Hybrid Cloud Computing Environments”, 3rd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2011), pp. 432-439, 29/11-01/12, Athens, Greece, (IEEE Best student paper award).
- P. Makris, D. N. Skoutas and C. Skianis, “Efficient Integration of LTE Femtocells in IP Networking Infrastructures: a Context-Aware Backhaul Link Exploitation Study”, POSTER paper presented in Femto Winter School organized by FP7 ICT IP BeFemto and STREP FREEDOM projects, held in Barcelona, 6-10 February 2012.
- P. Makris, D. N. Skoutas and C. Skianis, “On Networking and Computing Environments' Integration: A Novel Mobile Cloud Resources Provisioning Approach”, IEEE International Conference on Telecommunications and Multimedia (TEMU 2012), pp. 71-76, 30/07-01/08, Crete, Greece.
- N. Nomikos, P. Makris, D. N. Skoutas, D. Vouyioukas and C. Skianis, “A Cooperation Framework for LTE Femtocells’ Efficient Integration in Cellular Infrastructures Based on Femto Relay Concept”, 17th IEEE International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD 2012), pp. 318-322, 17-19/09, Barcelona, Spain.
- P. Makris, D. N. Skoutas, N. Nomikos, D. Vouyioukas and C. Skianis, “A Context-Aware Backhaul Management Solution for combined H2H and M2M traffic”, IEEE International Conference on Computer, Information and Telecommunication Systems (CITS 2013), 7-8 May, Piraeus, Greece.

1.6.3 Other Publications

There also are some other publications, whose material has not been used for this PhD thesis but are related to the extended research field of context aware resource management for mobile and fixed networking systems:

- P. Makris and C. Skianis, "Multi-Scenario Based Call Admission Control for Coexisting Heterogeneous Wireless Technologies", IEEE Global Telecommunications Conference (GLOBECOM 2008), pp. 1-5, 30/11-04/12 New Orleans, USA.
- P. Makris, G. Lampropoulos, D. N. Skoutas and C. Skianis, "New Directions and Challenges for MIH Operation Towards Real Market Applicability", 1st International Workshop on Mobility in Future Internet (MiFI '10), pp. 306-313, Chania Crete, Greece, July 2010.
- N. Nomikos, D. Vouyioukas, T. Charalambous, I. Krikidis, P. Makris, D. N. Skoutas, M. Johansson and C. Skianis, "Joint Relay-Pair Selection for Buffer-Aided Successive Opportunistic Relaying", accepted for publication in Transactions on Emerging Telecommunications Technologies (ETT), Special Issue on High Performance Mobile Opportunistic Systems, July 2013.
- N. Nomikos, P. Makris, D. Vouyioukas, D. N. Skoutas and C. Skianis, "Distributed Joint Relay-Pair Selection for Buffer-Aided Successive Opportunistic Relaying", 18th IEEE International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD 2013), Berlin, Germany, September 2013.

1.7 PhD Thesis Roadmap

The remainder of the thesis is structured in five chapters. Chapter 2 provides a literature review and related state-of-the-art works. In chapters 3-5, architectural and algorithmic innovations are proposed regarding context aware resource management: a) in a 4G HetNet environment (chapter 3), b) for mobile and fixed networking systems' convergence (chapter 4) and c) in mobile/hybrid cloud infrastructures (chapter 5). Finally, chapter 6 concludes the thesis' concepts and sententiously provides related future research directions.

CHAPTER 2

LITERATURE REVIEW AND CONTEXT LIFE CYCLE

In this chapter, a thorough literature review is provided based on the technical scope of this thesis research field as this has already been defined at a high level in chapter 1. More specifically, we initially define the Context Aware Mobile and Wireless Networking (CAMoWiN) area, we identify CA functionalities as CAMoWiN puzzle pieces and we end up by providing state-of-the-art work on CA resource management frameworks/techniques and ongoing research on mobile and fixed networking systems' convergence towards networking and computing environments' integration.

2.1 Context Aware Mobile and Wireless Networking Evolution

Context is a notion whose better understanding and use can enable the rapid evolution of CAMoWiN by providing insights into abstract mechanisms and functionalities required to support this field of research. Various past out-of-date and incomplete definitions have been proposed in the literature. Some early definitions in 1990s defined context by example [1] [35] and some others tried to interrelate it with other notions such as the environment or situation [36] [37] limiting thus context's applicability range. Dey in [38] achieved to provide a definition, which proved to be adequately diachronic for the initial stages of CA computing. According to [38], "context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves". This definition, apart from being extremely general, has some substantial drawbacks. Table 2.1 summarizes them providing thus five rationale points that differentiate past and up-to-date context definitions in the literature. First of all, nowadays, context is a collection of measured and inferred knowledge [39] rather than just a set of values with no underlying understanding of what these values ultimately mean. Secondly, there is a lack of clear separation between the concepts of context and context information [40]. The latter concept implies a process of exploiting the context information in various abstract ways and a perpetual flow of context information among local and/or remote heterogeneous entities. Thirdly, context does not simply "characterize the situation of an entity". Contrariwise, context can also arise from the general activity of a next generation CA system, thus generating and sustaining the context [41]. Another drawback of Dey's definition is that it takes for granted the fact that context exists only when "an interaction between a user and an application" occurs. This is not true especially in the field of wireless networking where for example if nothing is sent to a mobile node for a predefined time interval, the node can infer the context and proceed its

functionality without even making complex calculations. Finally, borrowing rationale points from autonomic networking, transparent management should be supported to users in CAMoWiN and thus "interaction" should not have to be explicit and noticed by the end users [42].

Table 2.1: Past vs. Up-to-date Context Definition rationale points

Past Context Definitions	Up-to-date Context Definitions
Context as a set of numerical values	Context as measured and inferred knowledge
Context as a state of information	Context as a flow of information
Characterizes the situation of an entity	Arises from the general activity of the CA system
Context as outcome of interactions	Context can exist independently of interactions
Users take part in system adaptation procedures	System adaptations unnoticed by users

As it has already been implied in the introduction, uncertainty management appears to be the strongest link between wireless networking and context awareness. As a result, we have to thoroughly investigate all possible aspects that uncertainty can appear in a general CA system, in order to effectively understand and define CAMoWiN area. Therefore, in figure 2.1, a classification of uncertain context information is provided. We claim that uncertain context information exists when: a) there is no clear knowledge of something (imperfection), b) it is difficult to distinguish an option among some seemingly correct alternatives (ambiguity), c) there is a mismatch between the actual and the reported states of information (wrong context), and d) no information is available at specific time instances (unknown). Various combinations of these four classes of uncertainty are possible (e.g. imperfect information can also become wrong under specific circumstances) and some attributes of one basic class can easily become attributes of another one, if specific use cases are assumed (e.g. out-of-date information can become inconsistent, if the system does not eliminate it on time). Below, we give a representative set of abstract examples and use cases regarding the various types of uncertain context information.

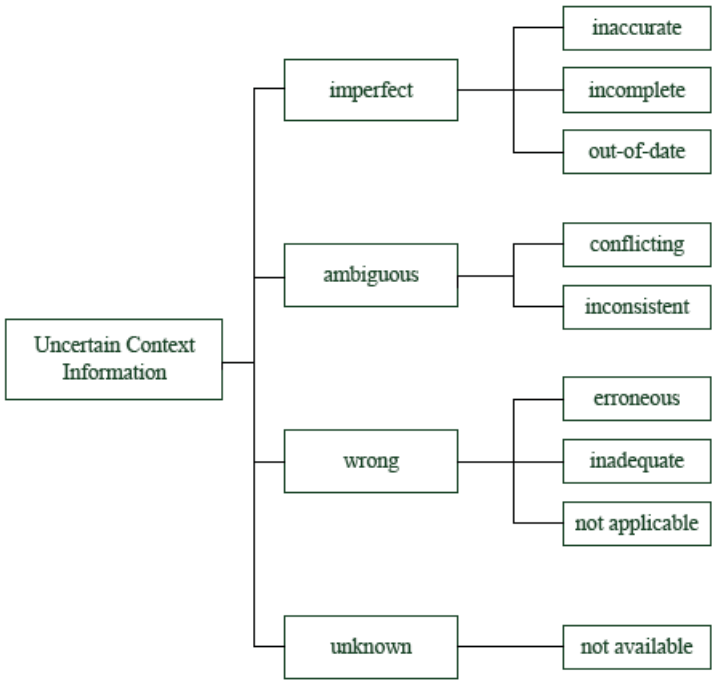

Figure 2.1: Classification of uncertain context information

Imperfect context is information or knowledge, which is basically correct but it has some shortcomings that are usually manageable by the system. These shortcomings appear in terms of inaccuracy, incompleteness and out-of-dateness. For example, when signaling messages are exchanged between a wireless access network and a mobile terminal (MT), possible incomplete context information will enforce the latter to guess the missing information in order to proceed with the appropriate procedures in case an immediate action has to be taken. If this can't be done, the MT will request once again for the same information. A common use case regarding inaccuracy emerges during collection of raw sensor data. Inaccuracy problems also appear during modeling and interpretation processes of context information and will be extensively analyzed in section 2.3.2. Out-of-date data is not usually a major problem, when context often changes. The system assumes a previous set of context information, which most of the times does not affect vital system operations at the expense of a sub-optimal solution. On the other hand, in cases where context is relatively static, out-of-date data may mean anachronistic information inappropriate for use and exploitation.

Ambiguous context is information or knowledge, which exists in a CA system in various facets making thus its direct exploitation a difficult problem to solve. The definition of context information using natural language is a traditional example of ambiguity, because the representation of information is very abstract and difficult to relate to the real world, hampering thus users' interpretation capability (cf. section 2.3.2). Moreover, in an environment where a number of policies need to coexist, there is always the likelihood that several policies will be in conflict, either because of a specification error or because of

application-specific constraints (cf. section 2.3.4). In [[43], a four-fold mechanism is proposed in order to deal with such conflicting issues, proposing thus identification, classification, detection and finally resolution of arising conflicts. Even though ambiguous context information can be resolved at early stages using conflict resolution techniques, inconsistencies (i.e. existence of information that contradicts or presents situations that severely violate established operating strategies) may also appear [44]. For example, a vertical handover (VHO) decision entity implemented in a heterogeneous wireless network environment has to select the optimum network among a list of competitors. In this use case, a conflicting issue may arise when a candidate network is the optimum choice for an end user but not a good one for the overall system performance in terms of utilization. This problem will be further processed and directly resolved leading to a conciliative sub-optimal solution. On the other hand, an inconsistency may be incurred if plenty of such sub-optimal decisions are made in a short time interval leading thus to violations of high-level and fixed strategies, whose existence ensure the long-term system's stability.

Existence of wrong context information is not an unusual phenomenon in a CAMoWiN system because much of the context being exchanged among heterogeneous entities is delivered via unreliable wireless links. Referring to early CA systems, erroneous context information arises as a result of human error and the use of brittle heuristics to derive high-level implications from low-level data [45]. In CAMoWiN systems, erroneous context may also be acquired by remote entities, thus not permitting MTs to process the received context. Wrong context can also appear in terms of inadequacy of information. One may observe that inadequacy is a synonym of incompleteness. Nevertheless, inadequate context means that the information lacks of some crucial elements, which cannot be recovered or inferred by the system ultimately classifying it as wrong. Furthermore, let us assume the same use case described for incomplete context information earlier in this section (i.e. when signalling messages are exchanged between a wireless access network and a MT). The sense of inadequacy lies in the fact that there is no available time for the MT to request for the same information because this would become out-of-date and generally useless (e.g. strictly real-time procedures). Not applicable context is the information, which is received by entities that are not able or should not exploit it for reasons of overall system efficiency. In CAMoWiN systems, large amounts of heterogeneous context is delivered in the wireless medium and thus mobile devices usually receive and process useless or practically wrong messages. This problem becomes even greater in next generation mobile and wireless networking where cooperative communications field emerges as a new research trend (read more in section 2.3.3 and femtocell as a relay concepts in the next chapter).

Finally, the meaning of unknown context is quite easy to understand. In traditional CA systems, unknowns usually result from sensor failures and various connectivity problems

[45]. On the other hand, in CAMoWiN systems, an extended concept is adopted considering unknown context as information, which is (in whatever way) unavailable at specific time instances. Therefore, in some cases, even though specific context may exist somewhere in the system, it has to be available the exact time it is needed for use, otherwise it may be considered as completely unknown by the system.

So far, by investigating all possible aspects that uncertainty can appear in a general CA system, we gave the first hints regarding the main requirements, which are necessary for managing the complicated and inherently uncertain structures of context. Moreover, we showed that the semantic notion and use of context has rapidly changed during the last decade following the evolution steps of mobile and wireless networking. In the rest of this section, we further show how context awareness (CA) meets some emerging next-generation networking and computing trends creating thus the need of CAMoWiN research area to evolve (see figure 2.2).

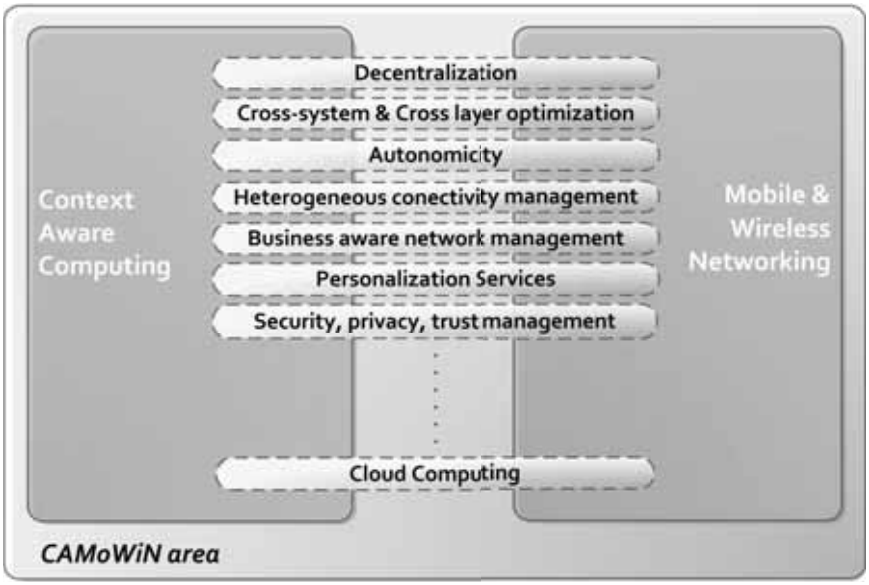

Figure 2.2: Cross-cutting challenges towards CAMoWiN evolution

One of the main challenges of next-generation networking is the lack of centralized goals and control [14]. Therefore, all recent research efforts propose either distributed system architectures or hybrid ones, where centralized components act as a subordinate system supplement [15]. Nowadays, various past shortcomings encompassed in hardware restrictions of MTs such as power supply, bandwidth, processing capabilities and storage are assumed to be manageable with the use of next-generation cognitive devices and thus distributed approaches can be implemented more easily [46]. Another major advantage of decentralized approaches is that they increase the robustness of personalized services and thus the invention of novel or killer-applications and services at a macro-scale is boosted. Therefore, today's CAMoWiN systems consist of a variety of distributed components, which have to be both

complicated and well implemented in order to be able to reduce the complexity of CA applications, improve maintainability and promote reuse [47]. Moreover, CAMoWiN systems adopt both cross-layer and cross-system optimization solutions, which have the ability to reliably establish a network-wide, global view of the overall system. Having such a global view, a MT can use global information for local decision processes in conjunction with a local view containing MT's-specific information contributed by each layer of the protocol stack or system component [48] [49].

Because of the fact that future networks are becoming increasingly dynamic, heterogeneous, less reliable and larger in scale, it is crucial to make them self-behaving, so that minimum human perception and intervention is needed in order to be managed and controlled [13]. Increasing context awareness in autonomic networks is a main cross-cutting challenge for CAMoWiN area [14]. More specifically, mobile and wireless networking research community is seeking ways to design a highly optimized system that will support distributed decision making and will incorporate in its architecture reconfigurable and cognitive aspects, in order to make it more autonomous. Similarly, CA computing community is seeking ways to design optimum context models that can uniformly drive autonomic decision making across the spectrum, allowing whole system self-optimization. Besides, as stated in [15], context awareness is a foundation of all self-x properties including self-configuration, self-organization, self-optimization, self-healing etc [16] (cf. section 2.3.4). Heterogeneous wireless connectivity management is a field where all these self-x functionalities can be applicable [17]. Furthermore, business-aware network management [18] (i.e. correlation of business objectives with network resources and services), CA security and privacy control [19] [20], personalized services [21], cloud computing [22] (in terms of efficiently provisioning right services for the contexts), phenomenological (in contrast with positivist) view of CA [41] and other are some of the latest concepts incorporated in the CAMoWiN paradigm.

2.2 Overview of Existing CAMoWiN Implementations

This section presents the related work found in the international literature regarding CA frameworks and architectures. In a nutshell, we initially refer some early efforts made for developing CA applications during 1990s and we go on by surveying the most popular CA middleware approaches, which have been proposed during 2000s. We also outline the ways some vital needs concerning context information's efficient handling and exploitation (apart from monitoring and gathering of context) came in the research foreground such as the need for exchanging context information with remote entities, the need for modeling and managing the context, the need to exploit context for optimal decision making etc. Furthermore, we

present latest state-of-the-art architectures and frameworks, which introduce innovative features incorporated in CAMoWiN (i.e. business logic, security and privacy control, energy/spectral efficiency, multi-disciplinary research, etc.) and can provide interesting envisions for the future. Finally, an accumulative table (i.e. table 2.2) containing a representative set of architectures and frameworks is provided accompanied by all main CA functionalities that can be supported by a CAMoWiN system. Therefore, the reader can visualize the gradual emergence and evolution of CAMoWiN area during the last two decades and can be better introduced to the issues being discussed in the remainder of this thesis.

2.2.1 Early CA implementations

The earliest focus of the research community regarding CA was how to collect, monitor and demonstrate the context in various CA applications (e.g. tour guides, shopping assistants, communication tools for mobile fieldworkers etc.). In [50], many of these early task specific CA applications are presented outlining that their major disadvantage was the fact that they suffered from tight coupling between the application and the underlying technology infrastructure (e.g. sensors) utilized to gather the context [51]. For example, in Cyberguide [52], when the sensors were changed, almost a complete rewriting of the application was required. Therefore, the use of sensor abstractions was introduced in order to solve the problem of inadequate sensor reuse. Moreover, additional functionalities such as limited storage and interpretation of context were adopted (e.g. Limbo system [53] and TEA architecture [54]). Dey's work was the first one who outlined the necessity of abstract frameworks and architectures in order to more easily design and develop widely applicable CA applications and systems. Context Toolkit provides a set of abstractions (i.e. widgets, interpreters, aggregators and discoverers) in order to facilitate the separation of application semantics from raw context data. Extended TEA architecture [55], extended Arch framework [56], Solar architecture [57] and WASP architecture [58] are some more example architectures dealing with the same issue.

2.2.2 CA Middleware Approaches

Encapsulating the context management logic into middleware rationale was another solution for the problem of decoupling context capturing and context processing from application composition [59]. Traditional middleware approaches emerged as a result of the need for more efficient representation and interpretation of context (e.g. Cooltown project [60]). There may also be multiple layers that context information goes through before reaching a CA application due to the need for additional abstraction and thus multiple middleware implementations (between two layers of abstraction at a time) are required [46]. Next generation middleware approaches were even more capable of adapting to environment

changes and supporting the required level of Quality of Service (QoS). For example, in CoBrA [61], SOCAM [62] and CAMPS [63] architectures, ontology engineering concepts were adopted in order context inference, modeling and reasoning functionalities to be upgraded. All the pre-referred proposals, including CMF [64] and CASS [65], assumed centralized architectures and frameworks in order to overcome storage, processing and energy constraints of traditional mobile devices. However, decentralized approaches gradually emerged along with the evolution of cognitive MTs and the need for exchanging and disseminating context information between remote entities. The primary goal of middleware approaches in distributed wireless environments is to focus on providing suitable abstractions for dealing with heterogeneity and context dissemination. A representative example is MobiPADS [66] middleware system, which first introduced the idea that both centralized network infrastructure and MT context information have to be taken into account for a decision to be made. PACE middleware's vertical handover (VHO) use case found in [47], goes one step further describing ways that a CA VHO procedure could be accomplished among heterogeneous coexisting wireless network technologies.

2.2.3 Context-dependent autonomic networking approaches

Modern CA architectures should provide appropriate mechanisms to achieve a suitable balance between user control and CA software functionalities autonomy. That is, CA applications not only require middleware for distribution transparency of components, but mechanisms dealing with system reconfigurability and adaptation, too. As stated in section 2.1, one main cross-cutting issue for CAMoWiN is how CA autonomic networking can be realized. CONTEXT Project [67] was the very first research initiative trying to introduce ideas on how context-awareness, ontology engineering and autonomic networking concepts can coexist in an innovative, extensible and scalable knowledge platform. Moreover, policy-based network management (PBNM) [68] concepts have been exploited in CAMoWiN in the essence that a CA architecture should support functionalities for sensing context changes and using policies specific to the new context in order to provide feedback to decision making entities [13]. Several CA architectures concerning CAMoWiN support context-dependent decision making functionalities such as CAPP [69], CA3RM-Com [70], AISLE [71], E^3 [16] etc.

2.2.4 Latest state-of-the-art CA architectures

Remarkable research work has been done in designing CA vertical mobility management architectures during the last few years. Cross-layer and cross-system architectures were designed and implemented for 4G network systems mainly dealing with CA resource management challenges seen from an overall system perspective (see section 2.1). CARP

framework [72] proposed ways to proactively manage network resources in a 4G heterogeneous wireless environment. Tramcar [73] and CrossTalk [48] architectures first introduced the idea of exploiting global networking context information in order to optimally design and implement various CA functionalities. Working further on this perspective, several CA architectures include IEEE Media Independent Handover (MIH) framework [74] and/or IEEE P1900.4 functional architecture [75] such as CAMMS [76], HURRICANE [49] [77], EMIH [78] and the P1900.4-based architecture proposed in [79], while [80] demonstrates ways that CA systems can benefit from a generic machine learning framework. The idea that CA can be seen from a holistic interactional point of view (i.e. context viewed by the prism of activity) was first introduced in BAC framework [41]. Finally, up-to-date research trends propose the incorporation of: a) business logic (e.g FOCALE architecture [13]), b) context-driven content creation, adaptation and media delivery via personalization services provisioning (e.g. C-CAST [81], CASUP [21]) and c) innovative security policy frameworks (e.g. SemEUsE architecture [82]) into next-generation CAMoWiN systems.

Table 2.2: Representative past & state-of-the-art CAMoWiN implementations

	CA Functionalities / **CA Architectures Frameworks Applications**	**Collect Monitor**	**Predict Infer Learn**	**Model Represent Interpret Reason**	**Store Retrieve**	**Exchange Disseminate**	**Evaluate Decision making**	**Business Logic**	**Security Privacy Trust**
Early CA implementations	CyberGuide [52]	**P**	**-**	**-**	**L**	**-**	**-**	**-**	**-**
	Limbo [53]	**P**	**-**	**L**	**L**	**-**	**-**	**-**	**-**
	Context Toolkit [38]	√	**-**	**L**	**L**	**-**	**L**	**-**	**L**
CA middleware approaches	Cooltown [60]	**P**	**-**	**L**	**-**	**L**	**L**	**-**	**-**
	CoBrA [61]	**P**	**P**	**P**	**L**	**L**	**-**	**-**	**L**
	SOCAM [62]	**P**	**L**	√	**P**	**L**	**-**	**-**	**-**
	CASS [65]	√	**P**	**P**	**P**	**P**	**L**	-	-
	MobiPADS [66]	√	**L**	**P**	**P**	**P**	**P**	**-**	**-**
	PACE [47]	**P**	**-**	√	√	**P**	**L**	**-**	**L**
CA autonomic networking approaches	CAPP [69]	**P**	**-**	**P**	**L**	**P**	√	**-**	**-**
	CA3RM-Com [70]	√	**-**	√	**P**	**P**	√	**L**	**-**
	E^3 [16]	√	√	**L**	**P**	√	√	**L**	**-**
Latest CA architect...	Tramcar [73]	**P**	**P**	**L**	**P**	√	√	**L**	**-**
	CAMMS [76]	√	**P**	**L**	**P**	√	√	**L**	**P**

	HURRICANE [49]	√	**L**	**L**	√	√	√	√	**P**
	P1900.4-based [75]	√	**L**	**P**	√	√	√	**P**	**P**
	CASUP [21]	√	√	**P**	√	**L**	√	√	**-**
	FOCALE [13]	√	√	√	**P**	**L**	√	√	**L**
	C-CAST [81]	√	**L**	**P**	√	√	√	**P**	**L**
	SemEUsE [82]	√	**P**	√	**P**	**L**	√	√	√

√: Supported **P: Partially Supported** **L: Limited Support**

2.3 Context Aware functionalities as CAMoWiN puzzle pieces

As depicted in table 2.2, the various CA architectures and frameworks are evaluated according to the extent of support they provide for eight different CA functionalities. We further sum up these functionalities to six and a complete analysis is provided for each one of them in subsections 2.3.1-2.3.6. Here, our aim is to have a global view of all CA functionalities in a CAMoWiN system in order to be able to propose valid technical contributions (in terms of novel frameworks, schemes and algorithms) throughout chapters 3-5. As depicted in figure 2.3, these CA functionalities are: a) context acquisition, b) context modeling, c) context exchange, d) context evaluation, e) exploitation of context from business logic perspective and f) CA horizontal functionality regarding energy efficiency, security, privacy and trust issues. More specifically, the context information flow is as follows: after raw context data is acquired from context sources and is appropriately processed, it is then transformed into low-level context, which serves as input to context modeling functionality. The outcome of context modeling (higher-level context) can be exchanged among various CAMoWiN heterogeneous entities before it serves as input to context evaluation functionality. The various final CAMoWiN system actions generated as output from various CAMoWiN decision-making entities can provide strategic business decisions according to human-oriented business logic principles. By the term “horizontal functionalities”, we mean all CA operations that can be applicable in any other CA functionality (e.g. energy, security, privacy, trust management mechanisms).

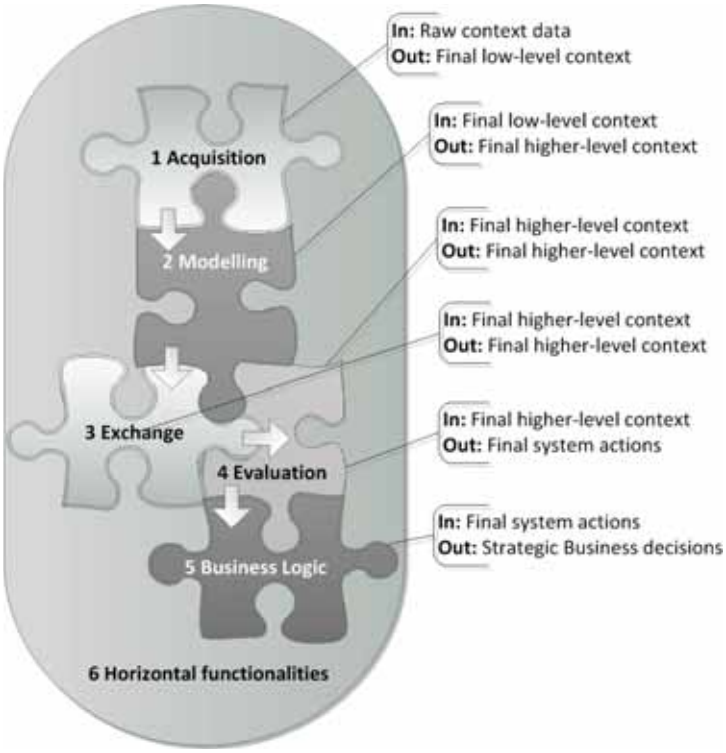

Figure 2.3: CA functionalities as CAMoWiN puzzle pieces

2.3.1 Context Acquisition

In general, raw data, modulated on any carrier (i.e. an electromagnetic wave), are transmitted by context sources via the wireless medium and are received by appropriate context acquisition modules. Following all the stages of context acquisition functionality described in this section, raw data are then transformed into low-level context information serving as input for the modeling functionality (see section 2.3.2). As stated in [83], context acquisition approaches should satisfy three main requirements: a) easy to deploy, b) easy to use, and c) non-intrusive for the end users. Taking all these into consideration and having made a thorough investigation in the international literature, we propose four main sub-functionalities of a CAMoWiN architecture regarding context acquisition: a) monitoring, b) gathering, c) predicting, and d) learning the context. In the rest of this section, the four sub-functionalities are further broken into distinct operations and are sententiously analyzed, following the structure of the taxonomy scheme depicted in figure 2.4.

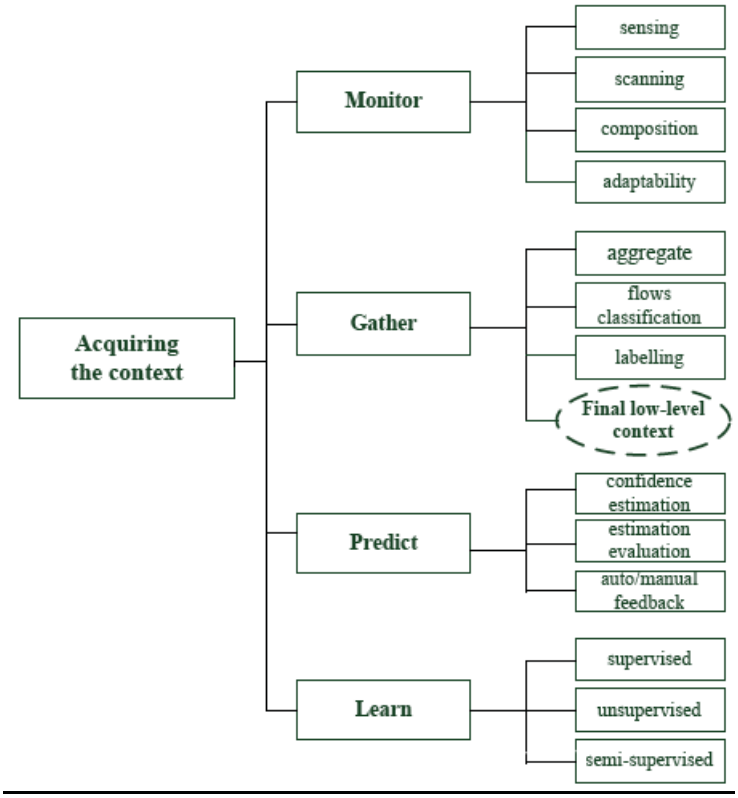

Figure 2.4: Context Acquisition Taxonomy scheme

In order to provide CA services to mobile users, there must be specific entities, which have to continuously monitor all types of contexts and capture their changes over time. A monitoring framework should deal with a broad spectrum of contexts, from personal to worldwide in terms of scale, and from crude to highly processed in terms of complexity. For example, in [84], context is classified in three main layers regarding personal, regional and global scale while in [85], it is clearly stated that a context monitoring framework should have the capability of monitoring physical to application-layer context information via appropriate cross-layer mechanisms. According to figure 2.4, the first operation of monitoring is the sensing of context, meaning both the detection and reception of data from numerous sensors residing in a specified CAMoWiN environment without introducing additional traffic overhead. In contrast to sensing, active monitoring approaches need to be provisioned in a CAMoWiN monitoring framework, too. Efficient scanning is a matter of vital importance in CAMoWiN, since real-time limitations have to be satisfied, disruptions of running services should be avoided and users' QoE objectives should be maintained within acceptable levels [86]. Moreover, scanning includes device and service discovery in order for the optimum set of context sources to be monitored and thus, to maintain an equilibrium between the needed accurate view of the CAMoWiN area and the processing overhead [87]. Following this principle, an optimum selection of multi-level contexts has to be made during composition phase. Finally, the "only when needed" property has also to be satisfied regarding the

monitoring adaptability operation. The key objective is the dynamic adjustment of the optimum set of parameters according to the use case being assumed by adopting novel event-based and on demand monitoring policies [88].

Following the discussion on the raw context data selection during the monitoring phase, gathering and pre-processing mechanisms are then considered in order to conclude with the final low-level context information (see outlined area in figure 2.4). A single mobile device or sensor is generally not capable (due to processing capacity, memory and battery limitations as well as due to the inherent heterogeneity and crudeness of raw context) to process the information facing it as a one-step procedure. More specifically, raw context data have to be aggregated, classified and labeled. During the aggregation phase, device readings are mapped into feature values such as mean or variance, using statistical and signal processing techniques, in order to reduce computational and communication overhead. In addition, aggregating algorithms recognize similar sets of data in order to merge them and/or eliminate redundancies [89]. After filtering all the unnecessary information, the classification task is responsible to find common patterns in the feature values. These so called classes or clusters represent initial semantic states such as “high temperature” or “crowded place” regarding sensors indicating weather conditions and population density correspondingly. The labeling task assigns descriptive names to combinations of classes in order for the low-level context information to be better interpreted and modeled during the modeling phase.

During the entire gathering procedure and especially for classification and labeling tasks, the performance evaluation of the algorithms being implemented is estimated on the basis of predictions delivered by trained models because of the uncertain nature of context information [90]. Prediction algorithms, regarding CAMoWiN acquisition frameworks, are usually applied to the classified context data and their outputs are of major importance since possible erroneous low-level context will definitely incur bigger problems in higher context abstraction levels. An overview of the most important prediction algorithms is given in [91]. Confidence estimation is a task of major importance for every prediction algorithm because the dynamics that govern the changes of lower-level contextual variables are chaotic and thus highly unpredictable. Therefore, each algorithm has to compute an estimation of the correctness of the forecasted context, which can be used as metric for the estimation evaluation task as well as for higher-level functionalities and decision-making entities (cf. section 2.3.4). Finally, a CAMoWiN prediction framework should support automatic as well as manual feedback of context information. That is, future context has to be compared with real context when this becomes available in future time series and consequently provide input to confidence estimation and estimation evaluation modules. At the manual feedback case, whenever the user cancels abruptly some action that has been carried out automatically due to a forecast, this fact is taken into account for future estimation and evaluation tasks [92].

Finally, both learning and adaptation to the users' behavior and their surrounding environment is necessary for context prediction, as a decrease in learning accuracy severely affects the context prediction quality [91]. Supervised learning is based on training data and is used in cases where non-real time data need to be processed [93]. In unsupervised learning, no labeled data are available for training and thus is employed in cases where quick and sub-optimal solutions are required [94]. Finally, semi-supervised learning allows taking advantage of the strengths of both by employing flexible use case-specific mechanisms according to higher-level policies ensuring thus CAMoWiN system's robustness and simultaneously utilizing system resources in the most efficient manner (cf. section 2.3.4) [95].

2.3.2 Context modeling

Final low-level context needs to be further processed in a CAMoWiN system. Context modeling is a functionality that interprets low-level context into higher-level context, reasons it to derive further implications and represents it in ways that efficient storage, retrieval and prediction/learning mechanisms can be applied. This functionality's output, which is the final higher-level context (see outlined area in figure 2.5), has to be in a standard format so that both context exchange and evaluation functionalities can efficiently exploit it as their input for their own processes (see more in sections 2.3.3 and 2.3.4). Requirements set for context models being adopted by CAMoWiN systems have been investigated in works such as [96] [97] and can be categorized in: a) easy manipulation on many devices supporting heterogeneity and mobility, b) providing low overhead in keeping models up-to-date, c) easy extension and reusability of modeling frameworks, d) efficient uncertainty management, e) timeliness, f) semantic reasoning, and g) scalability.

As shown in figure 2.5, in order to obtain high-level context information, the final low-level contexts (i.e. the output of context acquisition functionality) have to be further processed in the context interpretation step. The first task of the interpretation phase is the contexts aggregation. Low-level context data "carry" a huge amount of information that has to be filtered in order to be manageable by a CAMoWiN system regarding modeling functionality. Therefore, aggregation and fusion of related contexts, data calibration, noise removal and reforming of data distributions are some sub-tasks that are applicable [91]. Another main objective of context interpretation sub-functionality is to provide a common level of interpreted context going to be reasoned. This is not a trivial task as context information models deal with a large variety of context sources that differ in their update rate and their semantic level [96]. For example, context obtained from a CAMoWiN wireless sensor system is often heterogeneous and thus a common level of interpretation has to be applied in order system's consistency and robustness to be achieved. Finally, an additional categorization of context is required whose outcome will serve as input to reasoning sub-functionality. Hence,

the interpreted contexts can be categorized in sets each one of them providing identification, activity, spatial, temporal information etc [2].

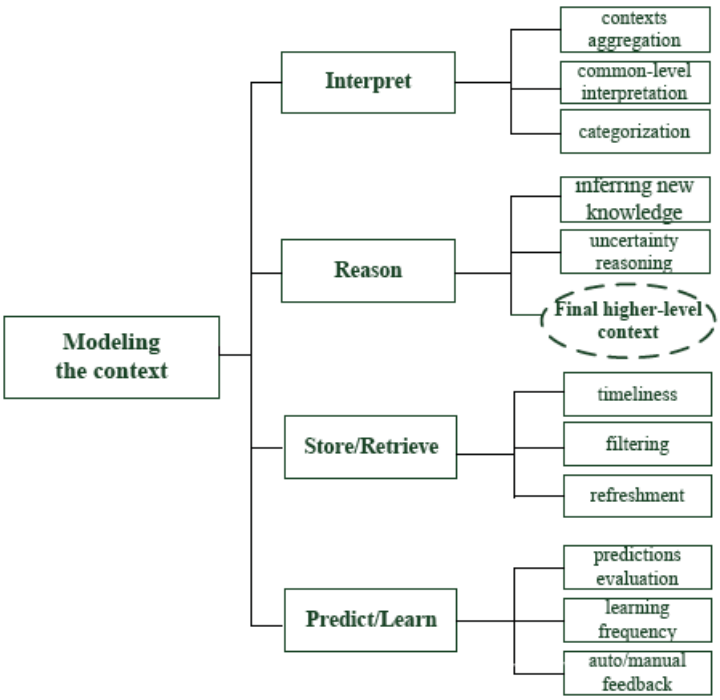

Figure 2.5: Context Modeling Taxonomy scheme

Context modeling can be applied in facts that are: a) observed or measured and b) inferred. In the former case, reasoning process can be eliminated whereas in the latter one, facts that are inferred can only exist by using reasoning to generate them [18]. Hence, reasoning refers to information that can be inferred from analyzing data and combining heterogeneous context information [97]. Context reasoning sub-functionality consists of two main tasks, namely a) new knowledge inference and b) uncertainty reasoning. Both of them refer to context information quality improvements, which can for instance be achieved by correctly recognizing and classifying the four types of uncertain context information described previously (figure 2.1).

Reasoning operations cooperate with the other modeling sub-functionalities such as storage/retrieval tasks for better uncertainty reasoning and prediction/learning tasks for better inference outcomes. Generally, storing all context information in CAMoWiN systems is impossible and technically unacceptable in terms of retrieval efficiency [64]. Indeed, the higher the level of the recorded data, the less data typically needs to be stored, but the higher the effort for retrieval [92]. A classification framework for storage and retrieval mechanisms is proposed in [98] from where we selected three main tasks, namely timeliness, filtering and refreshment, to present for the scope of this thesis. Timeliness refers to cognitive operations that decide whether a CAMoWiN system can reply to specific users' or subordinate system components' requests in specific time intervals. Hence, sub-optimal on-time solutions are

preferred compared to optimal out-of-date system responses. Filtering of context being stored and retrieved in large-scale CAMoWiN environments has become a necessity as the number of updates is very high and the quality of context has to be adequately good. Finally, refreshment deals with issues of eliminating and purging outdated stored context information, in order for the best possible context (in terms of both quality and quantity) to be available for retrieval and use at any time instance.

Regarding modeling functionality, context prediction and learning modules have to exchange information with the corresponding ones that process lower-level context data. Generally, lower abstraction level contributes to lower error probabilities and higher information content whereas higher abstraction level outperforms in terms of memory and processing requirements, which is equally important in CAMoWiN environments. Lower-level predictions evaluation, the learning frequency being applied and both the automatic and the manual feedback of context information required for reconfiguration procedures, are the three main operations of this sub-functionality. Generally, predicting unreliable context information may lead to catastrophic results. Therefore, predictions' evaluation task receives feedback from confidence estimation and estimation evaluation tasks described in section 2.3.1. Hence, only contexts, which are above specified thresholds in terms of uncertainty metrics, can be further processed in higher-level prediction and learning modules. Regarding learning sub-functionality, non-disruptive development of context models in CAMoWiN systems requires efficient and dynamic selection of supervised, unsupervised and semi-supervised learning algorithms according to the scenario being assumed [99] (see also section 2.3.1). Moreover, learning frequency is a task that is used to determine the time instances that a context database has to be scanned in order new prediction rules to be derived. Higher learning frequency means more up-to-date and thus more accurate prediction rules at the expense of higher processing overheads [91]. Finally, following the rationale of considering context prediction as a cross-layer functionality, auto/manual feedback task cooperates with the corresponding task of context acquisition functionality to achieve optimum reasoning of higher-level context information.

2.3.3 Context Exchange

In general, context information that is not being exchanged is useless for CAMoWiN environments. Given the fact that context information has been acquired and modeled in a CAMoWiN's system entity (sections 2.3.1 & 2.3.2), the next step is to be appropriately disseminated to more entities, which are interested in receiving and further processing the delivered contexts. The prerequisites that a context exchange framework has to satisfy have been investigated in works such as [100] [101] and the most important are: a) scalability, b) interoperability, c) adaptability, d) dealing with context uncertainties, e) content-based

dissemination, f) application-independent context exchange, and g) light-weight protocols' implementations. According to figure's 2.6 structure, the main CA functionality is split in three sub-functionalities as we want to a) define and categorize all possible entities (i.e. "who") that are participating in the emerging CAMoWiN ecosystem [102], b) define the parameters used for deciding "what" has to be exchanged at any time instance among the various actors discussed at (a), and c) define and categorize the means (i.e. "how") by which efficient end-to-end CA communication can be achieved.

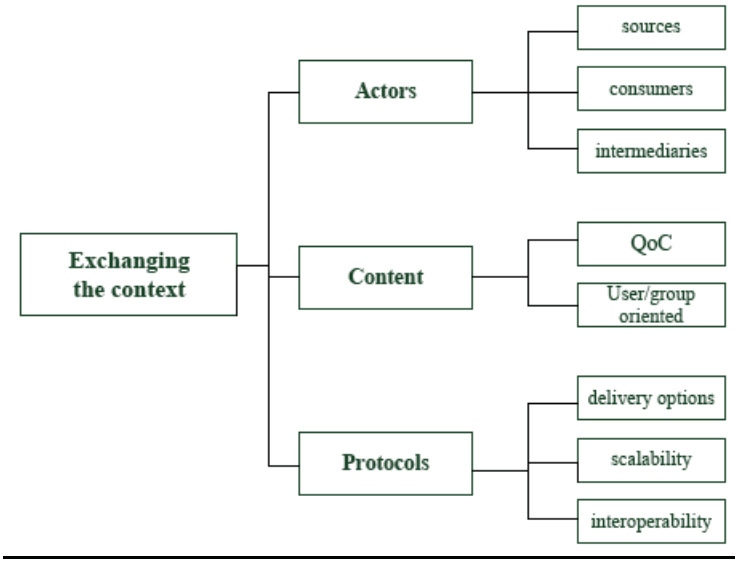

Figure 2.6: Context Exchange Taxonomy scheme

The term "actors" refers to all functional entities of a CAMoWiN system, which are interacting with each other in order to exchange context information and thus be able to perform optimal decision-making (see section 2.3.4). State-of-the-art research dealing with CAMoWiN actors [101] [102] [103] [104], categorize them in three main sets named context sources, context consumers and context intermediaries. Context sources refer to all CAMoWiN entities, which can generate and provide context information (e.g. location information providers, social networking sites, wireless sensor networks, etc). Based on the source of context information, sensor-equipped mobile terminals, telecom operators, sensor network operators, Internet platform operators, ICT companies, individual middleware developers etc. can be classified in one or more of the pre-referred categories. Context consumers are the final receivers, users and/or evaluators of the context information being exchanged [103]. In a future CAMoWiN environment, it will be common for an entity to act as context source for a CA application at a specific time interval and as context consumer for another CA application at a subsequent time interval. Given the basic provider-consumer model, only specific-purpose CAMoWiN systems could work with direct communication between context sources and consumers. However, in future large-scale CAMoWiN

environments, the existence of some kind of intermediaries or the so called context brokers becomes a necessity in order transactions costs to be reduced as well as scalability and interoperability challenges to become manageable.

Context-based content delivery is a research challenge, which deals with ways to provide personalized content to context consumers taking into account all contextual heterogeneities of a CAMoWiN environment. Ultimately, context exchange functionality has to be content-based instead of channel or subject-based [100]. State-of-the-art CAMoWiN architectures require good quality of context (QoC) to be exchanged in order to achieve efficient content delivery (regarding overall system's perspective performance evaluation metrics) [105]. Moreover, user/group-oriented content delivery should be applied. That is, although the same content is to be sent to all group members, its delivery should be adapted for different sub-groups based on their specific context. Efficient sub-grouping mechanisms are proposed in works such as [105] [106] [107] [108], which use cross-layer contextual information (e.g. user's profile, MT capabilities, network congestion, available bandwidth, link quality and characteristics, MT location and speed, quality of received signal, available wireless access technologies, etc.) in order different adaptations to be applied for various members of a group, which receive the same content.

Regarding context exchange protocols, the research is focused on efficient ways to exchange the content among the various actors. For example, a sequence of signaling messages for context-based content delivery can differ in situations where: a) strict QoS constraints have to be satisfied, b) energy consumption has to be minimized, c) specific security and privacy levels have to be achieved and d) unnecessary data transmission and processing in terms of optimal system performance can be avoided [109]. Therefore, by knowing and predicting the context, plenty of delivery options can be available regarding quantitative, qualitative and sequential features of messages being exchanged and thus optimal system performance evaluation achievements can be realized (read more in section 2.3.4). Scalability term refers to ways that context exchange protocols can deal with continuously growing factors such as: a) number of actors, b) number of interactions in terms of message exchanging between the actors, c) area of interaction and d) time span (e.g. more context has to be available for longer periods) [84]. In [103], various techniques are proposed for removing explicit dependencies between interacting actors facilitating the design of scalable CAMoWiN systems. [107] investigates the "scalability vs. personalization trade-off" problem, suggesting optimal solutions that provide personalized CA services without downgrading scalability objectives. Interoperability challenges emerge because of the large scale of CAMoWiN environments, too. HURRICANE architecture [49] [110] implements IEEE 802.21 MIH protocol [111] in order to provide interoperable means of communication among heterogeneous wireless network technologies, while P1900.4-based architecture [112] provides interoperable means

of communication among abstract MT and core network entities. [104] proposes an over-the-top implementation of 802.21 MIH protocol in order interoperability to be realized in mobile cloud computing paradigms, while [102] investigates possible business roles that actors will have in future CAMoWiN ecosystems.

2.3.4 Context Evaluation

Summarizing the previous three CA functionalities, we have come to a state where all actors of a CAMoWiN system possess all the higher-level context information they need in order to proceed with their final decisions and actions (see outlined area in figure 2.7). According to [64] [87], context evaluation functionality requires semantic richness of context information as input in order a CAMoWiN system to be able to change its behavior in response to a context change. In today's CAMoWiN environments where a vast number of cognitive mobile terminals, innovative mobile applications and different types of access networks are available, evaluation should be context-dependent and consider several factors, at different abstraction layers [17]. More specifically, a context evaluation scheme should at least satisfy the following general prerequisites: a) efficiently handle multiple, dynamically changing and potentially unexpected situations, b) guarantee optimum system's resources management, c) keep users' satisfaction levels above predefined thresholds providing personalized services [113], d) deal with complex optimization problems supporting real-time decision-making procedures [114] and e) minimize human intervention [115].

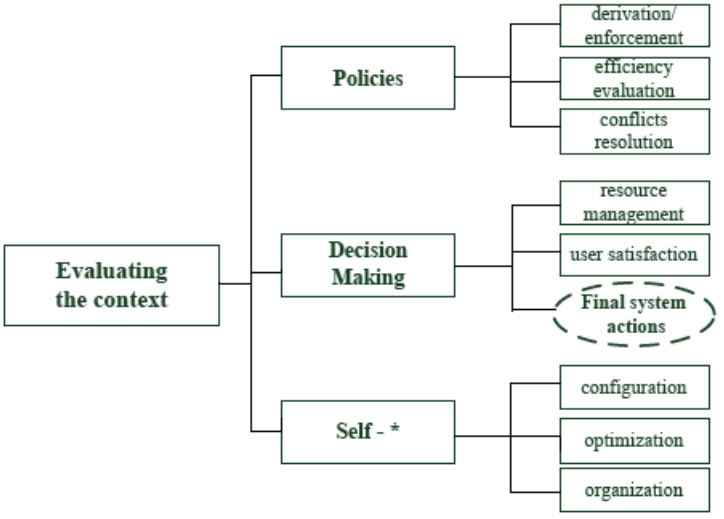

Figure 2.7: Context Evaluation Taxonomy scheme

Policies are used for abstracting specific actors' objectives at various levels such as technical, business and administrative. Policies can be applied to sets of resources to configure them uniformly and adapt their behavior to suite changing requirements, while possible policies' amendments lead to changes in goals (instead of specific system configurations) that are

implementation independent. There are three main abstract types of policies, namely: a) action policies that are based on event-condition-action rules, b) goal policies that specify a desired state, and c) utility policies, which express a value for each state to indicate how desirable it is [116]. There are several types of policies having being proposed in the literature, such as: a) generic resource management policies [79], b) radio access network selection policies [117], c) generic handover policies [110], d) dynamic spectrum assignment policies [114], e) self-organizing networking policies [16], f) scanning policies [86], g) energy-saving policies [118] and h) security policies [119]. In this thesis, we assess that future policy-based CAMoWiN evaluation frameworks should simultaneously take into consideration many of the pre-mentioned types of policies. Regardless of the applied type of policy, there are certain tasks that have to be supported, namely: a) policies derivation and enforcement, b) policies efficiency evaluation, and c) policies conflicts resolution. All these operations reside mainly in entities at the core network side, which have the capability to collect CA information on all MTs of a CAMoWiN system. The aim of a policy derivation module is to derive policies, whose goal is to provide MTs with guidelines towards making optimum decisions in order to ensure profitable operation for both MTs and overall network system's perspectives. The enforcement of the derived policies to each MT highly depends on the policies' grade of obligation (GoO); that is the degree of importance that each policy has according to its: a) purpose, b) imminence, c) impact, d) freshness, e) complexity etc [79]. Policy efficiency evaluation task aims at enhancing the knowledge acquisition capability of the policy derivation operation and thus improving the efficiency of future policy derivations. In a nutshell, this task should be cognitive enough to "understand" whether bad value indications happen due to: a) bad wireless channel conditions, b) bad information handling and decision making by MTs or c) the fact that current policies indeed have to change (out-of-date, conflicting, not applicable any more, etc.) in order the efficiency evaluation reports back to the policy derivation modules to have a high degree of effectiveness. Finally, conflicts resolution task aims at detecting and resolving existing and potential policies' conflicts and inconsistencies in CAMoWiN systems. The larger the number of policies and the more complex they are, the greater the likelihood of a conflict. Policies' conflicts importance degree can be estimated based on: a) the policy types involved and the scope of their enforcement (e.g. resource management policies may have greater impact factor compared to scanning policies), b) application environment constraints (e.g. a radio access network (RAN) selection policy may opt for a RAN that is not available at a given time instance), c) nature of their occurrence, and d) the time frame at which they can be detected (static vs. run-time conflicts) [43].

Decision-making frameworks and algorithms were usually not context-aware and did not take into consideration context information from several sources simultaneously. [120] is one of

the earliest works dealing with resource management issues in mobile and wireless networking. In 00's, when many new wireless network technologies had been standardized, novel challenges came up such as: a) heterogeneous environment, b) multiple types of services, c) adaptive and cognitive resources allocation, and d) cross-layer design exploiting context information from multiple abstraction layers [121]. Autonomic decision-making refers to the ability of a CAMoWiN system to perform adaptation operations using its internal knowledge to decide why, when, where and how adaptations are performed, without any human intervention in the decision making process [85]. [106] studies efficient ways to make decisions about delivering context information to specific context consumers' subgroups in order multiparty multicast session management procedures to be realized. [104] proposes ways on making intelligent decisions regarding network resources management in mobile cloud computing environments. Vertical handover (VHO) decision-making is also a "hot" topic in CAMoWiN field, while possible synergies between IEEE 802.21 [111] and IEEE P1900.4 [112] standards seem to provide good architectural solutions. Finally, works such as [122] and [123] investigate the distribution of decision-making functionality between MTs and the core network side trying to define key performance indicators regarding users' satisfaction levels.

Decision-making solutions can be categorized as: a) generic resource management mainly accomplishing operators' objectives and b) personalized decision-making for each individual MT mainly satisfying end user needs. Till recently, there was not a clear difference between the two pre-referred decision-making variants. For example, our proposed call admission control and scheduling algorithms [117] [124] try to keep some important metrics (delay, blocking/dropping probabilities, number of HOs, etc.) under predefined thresholds providing optimal trade-offs between conflicting operators' and end users' objectives. In today's CAMoWiN systems, user satisfaction does not only mean the accomplishment of some static average values milestones but the capability of each individual user to acquire personalized and specific QoS requirements' services each time he demands them independently of his previous or next demands and the demands of other users placed in the same geographical area. Conclusively, final system actions, which are the output of context evaluation functionality, comprise the highest-level context information serving as input for business logic functionality presented at the next section (i.e. 2.3.5).

Self-* (cf. figure 2.7) properties are adopted as a recursive operation of context evaluation functionality in order optimal decision-making procedures to be realized. Many variants of self-* properties have been used in the literature such as self-management, self-testing, self-protection, self-tuning, self-planning, self-healing, self-maintenance, self-stabilization etc. In this paper, we classify them in three main categories namely self-configuration, self-optimization and self-organization. Self-configurable is a CAMoWiN system, which can

detect changes happening in any of its actors or actors' interactions that can result in a violation of management objectives, and trigger appropriate (re)-configuration mechanisms without affecting the system's smooth operation. [125] is a representative example of self-configuration applicability in CAMoWiN field, where context knowledge (learning from past) can be exploited for real-time and cost-efficient network adaptation to continuously changing environment conditions. Given the fact that self-configurations are often sub-optimal, self-optimization modules are in charge of performing any context evaluation function in the most efficient manner. A self-optimized framework is proposed in [16], where optimal self-tuning of handover parameters based on corresponding HO policies, is investigated. Similar self-optimization paradigms can be applied for all types of policies mentioned previously in the current section. Finally, self-organization refers to dynamic (re)-organization of overall system structures for specific optimization goals. The main difference with self-configuration operation is that self-organization deals with global CAMoWiN system's objectives instead of configuring parameters at a component level [16].

2.3.5 Context exploitation from a business logic perspective

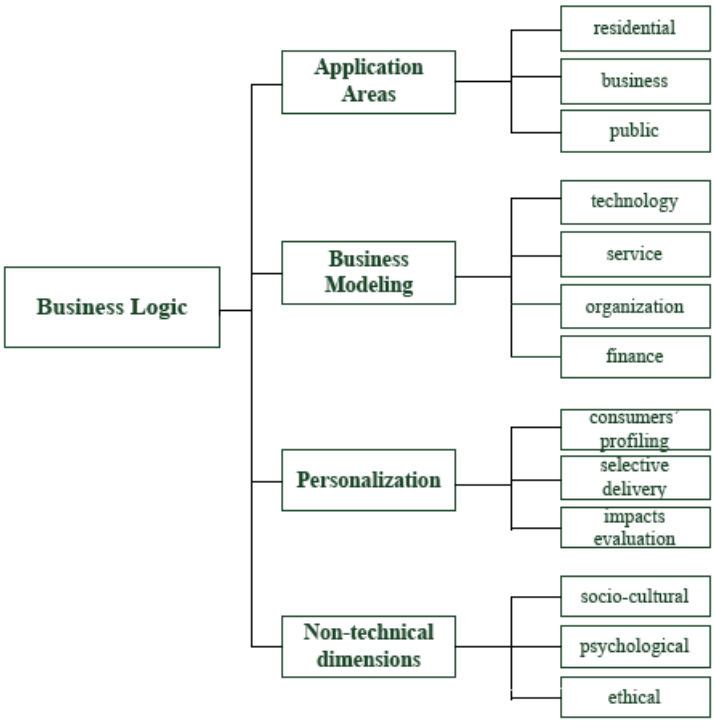

Figure 2.8: Business Logic Taxonomy scheme

During the last years, business logic challenges [15] gain on momentum in CAMoWiN field, since there is a tendency to focus human attention on the business logic rather than on any kind of configuration details already discussed in the previous four subsections. In figure 2.8, we classify all business logic functionalities assuming that final system actions (cf. figure 2.7)

are used as input for the whole CA functionality being presented in this section leading to the generation of strategic business decisions, which is the output of business logic functionality.

Given the fact that all CAMoWiN systems should adopt a business strategy, this strategy should take into consideration the corresponding geographical area, the environmental context and the mobile users' profiles. Hence, a first categorization regards to ways that different application areas can affect the applicability of diversified business strategies. Recently, several research proposals utilize various real market network deployment scenarios and use cases in order to better demonstrate the proof-of-concept of their approach [32] [113] [117]. We can recognize three main application area categories, namely residential, business and public CAMoWiN environments (see figure 2.8). Residential environments are associated with the intelligent home concept of a family or individuals living in rural, sub-urban or urban areas. Many different types of applications can be supported such as personal care, security and safety, electrical power management, entertainment and autonomous wireless communication. Generally, small number of very demanding (in terms of QoS and QoE) paying home users are assumed whose profiles are well known to the system but their demands for a large variety of services during the day can be highly unpredictable. The users themselves deploy and manage the CAMoWiN system and this entails the fact that they want to pay the minimum cost for the best QoE services they can get. From the operator's perspective, this brings on some technical challenges as each small-scale CAMoWiN system has to incur no problems to its neighbors (e.g. interference, QoE degradation, security) and simultaneously to optimize operator's revenues regarding the large-scale CAMoWiN ecosystem. Business environments include enterprises, institutions, organizations (e.g. ministries, other governmental buildings, etc), campuses, conference places etc. In this kind of environments, CAMoWiN systems serve both permanent and guest users, whose profiles are at a large extent known. Given the fact that the area, the users and the services are purpose-specific, the CAMoWiN system has to satisfy the corresponding users' needs without compromising the QoE for any other conventional service they enjoy independently of their stay at the purpose-specific geographical area. Finally, public environments can be metro/railway stations, airports, shopping malls, stadiums, crowded city squares/streets etc, where the widest possible set of diversified users, devices, services can be assumed. Thus, the main challenge for a CAMoWiN system is how to "understand" all the personalized needs of each individual and optimize QoE for all users at any time instance.

Regarding business models, they provide the means for the support of the application areas described above. Nowadays, business modeling, in terms of improving business models' viability and attracting more end users and enabling potential actors to deliver attractive solutions within the value chain, has become a very "hot" issue in the CAMoWiN industry [97] [126]. Regarding technological requirements that a CAMoWiN business model has to

satisfy, the most important are: a) balance between QoS experienced by mobile end users and corresponding costs, b) acceptable threshold achievement of what users allow others to know about their context (i.e. efficient users' profiles management), c) making optimal decisions for the "where to put intelligence of personalization" problem by using novel technological solutions (e.g. processing in the cloud [127], cooperative relaying [128], etc) and, d) balance between security and ease of use. Similarly, from the service point of view, some key requirements are: a) existence of multiple versions of a service to satisfy different priority groups, b) balance between complete context awareness at the expense of users' time and money, c) avoidance during the development of a new service, underestimations from a long-term perspective and overestimations from a short-term perspective, d) existence of trusted actors, and e) efficient and dynamic evaluation of the degree of personalization being offered. From organization perspective, we need to have: a) separation of roles (all actors should find their own niche), b) market value chain openness (existing operators should be open to provide CA information to new actors being compensated for this service), and c) proper governance via SLAs among highly competent actors. Finally, regarding financial features, we need: a) good pricing schemes (CA services should worth be paid to use and investigate how mobility offers adequate added-value to end users), b) deal with multiple revenue models (e.g. advertising, subscription, transaction-based), and c) balance between costs and revenues for all CAMoWiN system's actors.

Personalization is all about "understanding the needs of each individual and helping satisfy a goal that efficiently and knowledgeably addresses each individual's need in a given context" [129] given the fact that in today's CAMoWiN systems, the services that users want to receive are different despite of the same context [130]. Referring to personalization sub-functionality, three main operations can be identified, namely a) profiling of consumers, b) delivery of selective services and c) evaluation of appropriate impacts.

Non-technical dimensions and impacts of new technological paradigms are not usually given a significant consideration during development process. However, a major research topic that worth investigation is how can CAMoWiN's area sustainability be guaranteed regarding non-technical impacts such as social, cultural, psychological, legal and ethical issues (cf. figure 2.8) [133]. Seen from the socio-cultural point of view, sustainability of human society is based on certain levels of privacy, anonymity, trust and limited knowledge (see more in 2.3.6). For example, it is very doubtful that the society is adequately equipped with the right sociological and technical tools in order to understand what can happen when everything is available, knowable, searchable and recorded by everyone all the time (e.g. in social networking sites). From context sources perspective, it is very important to take into consideration the socio-cultural profile of context consumers in order to avoid useless delivery of context information that may even frustrate end users (for example imagine the

case of a regular employee getting a message about a special trendy cloth discount costing "only" 1000 euros!). Psychological aspects have to also be taken into account, since a wrong decision-making procedure by a CAMoWiN system can irritate end users at the extent of abruptly stopping the usage of a specific CA application for a long time. Minimizing intrusiveness is also important, since CAMoWiN systems should not give the impression to users of an outsider always monitoring them. It is also challenging to find optimal trade-offs between the usefulness of recommendations and the amount of privacy the consumer needs to give up to receive acceptable-quality personalized offerings as well as determining ways to minimize intrusiveness while maintaining personalization quality [131]. Finally, CAMoWiN systems should extend the human subconscious functionality by acting at the right time and by providing the right level of information. Simultaneously, all these can happen only if ethical and legal issues are by no means compromised. Conclusively, ICT and human values have to be integrated in such ways that the former enhances and protects the latter, rather than damaging them [134].

2.3.6 CA horizontal functionalities

In this section, a taxonomy scheme is provided for CA horizontal functionalities, namely: a) energy efficiency, b) security, c) privacy, and d) trust issues.

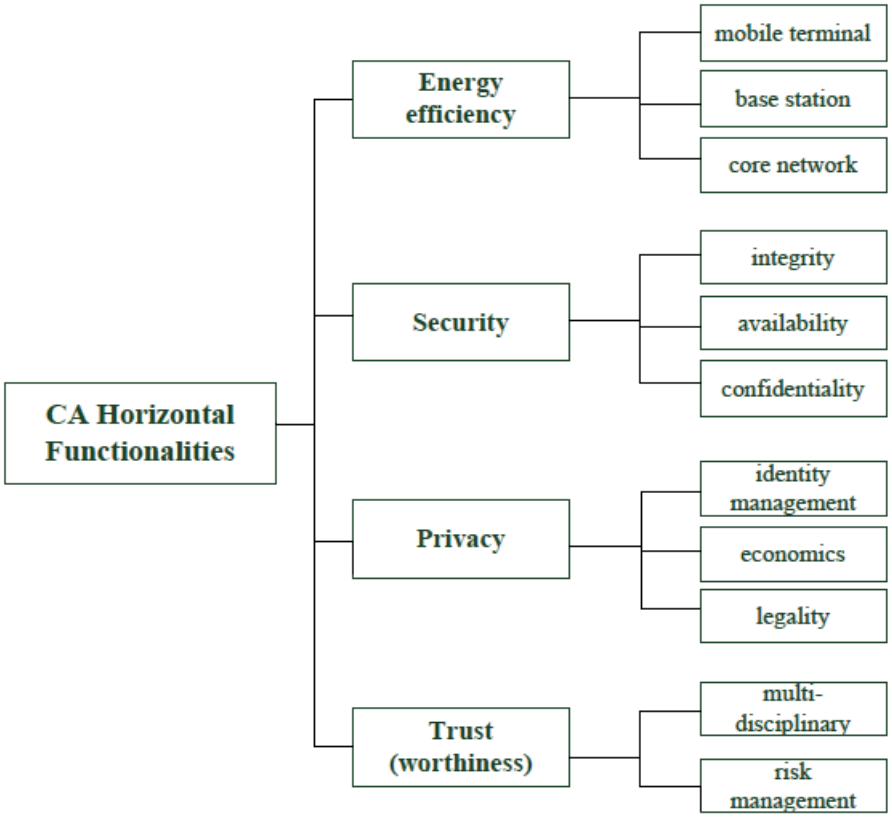

Figure 2.9: CA Horizontal Functionalities Taxonomy scheme

Energy efficiency (EE) in CAMoWiN area is a growing concern for cellular/wireless network operators in order to not only maintain profitability but also reduce the overall environmental effects. EE solutions have also practical impact for mobile devices, which want to save battery power in a seamless way and simultaneously not incur interference problems to the overall CAMoWiN system [135] [136]. Saltzer and Schroeder were the first who defined: a) the notion of security as set of mechanisms that control the use of context information, b) the notion of privacy as the ability of an entity to determine whether, when, and to whom information is to be released and c) the notion of trust, which denotes the grounds for confidence that a system will meet its security objectives [137]. Nowadays, citizens are increasingly concerned with the idea that their private data or personal profiles can be stolen or used for commercial purposes without their consent, that they will have to compromise on being permanently monitored by outsiders and that there exists a persistent risk of being electronically robbed or cheated by a stranger whom they have never met without understanding how this happened and with little chance for legal redress. Therefore, we envision CAMoWiN sustainability via investigating (security-privacy-trust) SPT problems is this section's case, because if CAMoWiN users feel threatened, mistrustful and increasingly hesitant towards using innovative CA mobile applications and services, then the ultimate loser will be our society as a whole [138]. Each one of the CA horizontal functionalities cannot be seen as a single and isolated functionality in CAMoWiN architectures but instead, provisions of CA horizontal functionalities should be considered within all five CAMoWiN functionalities (sections 2.3.1-2.3.5). Hence, we consider CA horizontal functionalities split in several modules assisting other CA functionalities rather than comprising a subordinate element in a whole CAMoWiN system (cf. figure 2.3).

EE issues in CAMoWiN can be classified in three main categories according to the part of the networking infrastructure the research innovations are applicable. Energy saving technologies have been proposed for multi-standard wireless/mobile terminals, exploiting the combination of cognitive radio and cooperative strategies, while still enabling the required performance in terms of data rate and QoS to support active CA applications [135]. Moreover, ways that context information can be used by cooperative strategies to achieve power efficiency at the wireless interfaces of mobile terminals and save battery lifetime are also being discussed. From the base stations' (BS) perspective, innovative techniques such as: a) cooperative BS power management, b) H/W enhancements, and c) renewable energy resources are studied [136]. EE techniques are also being proposed for fixed networking systems in the packet core network mainly dealing with: a) EE network design, and b) EE network operation [139].

Three main security variants needed for a CAMoWiN system are assumed: a) integrity, b) availability and c) confidentiality [140]. Integrity operation focuses on ensuring that CAMoWiN system's assets (i.e. hardware, software, media storage and data, etc.) can be

modified only by authorized entities and on guaranteeing that the provided context information has not been corrupted by any third party. Guaranteeing the prevention of unauthorized modification of data, hash functions and public/private key infrastructures can provide context information integrity. Digital signatures are also useful for future integrity testing needed for corruption recovery [141]. Availability operations focus on ensuring that CAMoWiN system's assets are available to the authorized entities and thus all information resources are accessible to legitimate users when required. Denial of Service (DoS) attacks are the common techniques to cause information resources of context information provider to become unavailable incurring sleep deprivation torture and CPU exhaustion effects. Conclusively, ensuring availability (for example) in e-health/e-government/e-banking applications means that all system's assets are expected to be online in order to prevent inconsistencies during patients'/citizens'/bank customers' requests [34] [140]. Finally, confidentiality operations provide the protection of sensitive information from unauthorized access ensuring thus that information is accessible only to those authorized to have access. Modern cryptography (i.e. symmetric and asymmetric encryption techniques) can provide good solutions; alternatively Cloud Computing vendors adopt physical isolation and virtualization approaches [141]. Fine-grained access control policy-based mechanisms, which determine whether an access request to a resource or data is granted or denied, are also applicable solutions [34].

Privacy is the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others [142]. There are six main requirements for a CAMoWiN system to support privacy [143]: a) notice (collection practices, privacy policies and policy announcements should be efficiently declared), b) choice & consent (users should have the choice of carrying out, or not, their personal data), c) anonymity & pseudonymity (whenever the users' identity is not required or whenever the user does not consent, anonymity or pseudonymity services should be provided), d) proximity & locality (the collection of data from a user's device should only occur when the user is present while processing and access to these data should only be done within the space of collection), e) access & recourse (access to the user's data should only be allowed to authorized persons) and f) adequate security (in terms of the security principles analyzed above). Privacy of personal data should be supported by identity management (i.e. before users release data) and trust management (i.e. after users release data). Specifically, identity management frameworks enable users to limit the amount of personal information revealed to certain providers or in certain applications. Another hot issue in the CAMoWiN market is the economics of privacy, which studies the trade-offs associated with the protection or revelation of personal information [126]. However, apart from technical and financial perspectives, privacy is primarily a human right and thus corresponding global law protection frameworks

should be addressed. Nowadays, the most complete and CAMoWiN-related law framework is the EU Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector (continuation of Directive 95/46/EC). After all, a good law framework being adopted by the majority of governments and institutions all over the world will assist CAMoWiN's area sustainability because context consumers are generally more open to disclosure when legal requirements are respected.

As stated above, privacy of personal data should be supported by trust management architectures, too. In general, trust is another mechanism to cope with uncertainty issues and consequently it is highly context dependent [144]. In CAMoWiN scope, trust refers to the reliability of the context information being delivered and can be evaluated by computing the distance between it and the real context. As a result, the term "trustworthiness" is used to represent the level of trust that can be assigned to one party (B) by another party (A) to do something (X) in a given relational context and thus is a measure of the objective probability that the trustees will behave as expected by the corresponding trustors [138]. Nowadays, it is well accepted that coordinated multi-disciplinary solutions should be adopted for trust-oriented CAMoWiN systems. Trust management is inherently related to risk management theory because the most important part of managing trust is understanding the risks involved in trust-based interactions [145].

2.4 State-of-the-art on Context Aware Resource Management

In this section, a thorough survey of all joint call admission control (JCAC) being applicable in heterogeneous wireless access networks is provided emphasizing in emerging context-aware enhancements recently proposed in the international literature. Moreover, we present all major existing capacity partitioning techniques, which are used for resource management problems in mobile and fixed networking systems. More specifically, in 2.4.1, complete sharing (CS) and complete partitioning (CP) techniques are presented, which are used as reference schemes for the schemes being proposed in the current PhD thesis (see more in chapters 3-5). In subsection 2.4.2, more recent approaches such as hybrid (HP) and virtual partitioning (VP) techniques are presented. Some of this thesis' proposed schemes follow the main concepts of HP and VP, while they are complemented with some novel advanced context-aware resource management approaches such as the ones provided in 2.4.3.

In general, JCAC schemes are needed in order radio resources to be jointly managed. JCAC is a major category of JRRM and is responsible for deciding whether an incoming call/service can be accepted or not and which of the available radio access networks (RANs) is/are most suitable to accommodate the incoming call/service. The decision strategy for the admission of new connections consists of guaranteeing that the sum of the minimum rate requirements of all accepted connections does not surpass specific thresholds. This strategy guarantees that

the resources available to a scheduling module are sufficient to provide the QoS requirements for all accepted connections. The major factors that need joint management from a JCAC perspective are: a) multiple heterogeneous RATs, b) multiple user groups with diversified priorities and needs, and c) multiple service groups with diversified QoS requirements. Furthermore, the major network key performance indicators that have to be taken into account and are the objectives of JCAC in mobile/wireless networking are [146]: a) guarantee QoS/QoE requirements (data rate, delay, jitter, PER, BER) of accepted calls/services, b) minimize call blocking and dropping probabilities, c) maximize operators' revenues, d) maximize radio resource utilization, e) maximize user satisfaction by granting additional resources beyond those required in the initial AC process, f) minimize the number of handoffs from one RAT to another, and g) uniform distribution and balancing of the total network load. From the above, it can be easily inferred that there are trade-offs when trying to satisfy some of the pre-referred objectives. As already mentioned in section 2.3.4, multi-objective decision making rationale can be applied in such kind of problems. Below, we further elaborate on various partitioning techniques ending up by describing the roadmap for future JCAC implementations being applicable in convergent mobile and fixed networking environments (e.g. 3GPP heterogeneous networks (HetNets), mobile cloud, etc). Partitioning refers to various algorithmic techniques that can be applied to a pool of resources trying to fulfill multiple and diversified objectives. In the following, complete sharing, complete partitioning, hybrid/virtual partitioning and advanced related context-aware approaches are described.

2.4.1 Complete Sharing and Complete Partitioning

Complete Sharing (CS) is the most trivial technique and considers one unique pool of resources, which is common for all combinations of user groups, service groups and available RATs. A new call/service is admitted into the system, if there are adequate resources from the unique common pool, otherwise it is rejected. That is, when the total network resources get to their limits, a new call will be blocked while a handoff call will be dropped. CS is a first-come-first-served (FCFS) non-prioritization scheme and thus adopts the simplest resource allocation policy. Its major advantages are implementation simplicity and high radio resource utilization. However, CS does not provide any QoS differentiation and thus has poor QoS performance [147].

In Complete Partitioning (CP), the overall resources are partitioned into several parts according to a combination of RAN, user group and service type and a new call/service request is rejected, if the resource mapped to the corresponding combination is used up [148]. In other words, by the term "partitioning", we mean that a fixed capacity is allocated to each combination and thus no resources from one partition can be allocated to more than one combination. The size of each partition is defined according to a priori knowledge that the

system already acquires taken from extensive past statistical measurements. From this kind of measurements, various mobility and load traffic patterns are derived and thus a good calculation of the size of each fixed partition can be done. Whenever a radical change in JCAC-related key performance indicators is observed, a network administrator can manually calibrate various parameters assuring the system's proper operation. The main advantage of CP is that it has good QoS performance and ensures the fairness of different priority calls, but due to the fixed partition policy, the radio resource utilization can be severely decreased. Another main drawback is that there is no elasticity/dynamicity in the size of the partitions and no intelligence is included in the JCAC process.

Towards providing more intelligence in both CS and CP schemes, many algorithmic proposals have been made in the literature during the last decade. "Guard channel" is one of the earliest techniques and it proposes to reserve some extra capacity for prioritized calls (e.g. handoff calls, high-priority user/service groups, etc) by implementing a static threshold. Higher utilization can be achieved through dynamic adaptation of the threshold according to the network state by adopting the "fractional guard channel" scheme [121] [149]. As the dynamic adaptation depends on the radio resource utilization, various acceptance probabilities can become smaller, when utilization is high and vice versa. Enhanced proposals based on fractional guard channel are multi-threshold resource reservation schemes, which implement multiple dynamic thresholds in order to assign different priorities to multiple combinations of service calls [117] [150]. Thinning algorithms [151] follow the same rationale by supporting multiple types of services and calculating the admission probability based on the priority and the current traffic situation. Finally, in queuing priority schemes, when utilization reaches 100%, high-priority calls are queued and are served when some radio resources become available. In this case, the queuing delay has to be inter-related with other types of delay imposed by scheduling process. The main drawback of queuing priority scheme is that is needs a lot of buffers to deal with real-time multimedia traffic. It also needs a sophisticated scheduling mechanism in order to meet the QoS requirements of delay-sensitive calls [146].

2.4.2 Hybrid and Virtual Partitioning

The problem with CS and CP schemes is that they are not flexible enough in order to cope with all emerging JCAC challenges (e.g. integrated services admission control, multi-homing, innovative mobile services/business models, etc). Nowadays, users' demands are not only restricted in enjoying different types of services from various heterogeneous RATs but they continuously and increasingly demand for more flexibility and elasticity in order their QoE to be enhanced. Innovative business models are also pushing towards this direction as an individual user may have more than one profile according to the mobile device he/she uses, his/her location, etc. As a result, the number of the corresponding combinations referred in

2.4.1 has become very large and conventional partitioning schemes cannot handle the incurred algorithmic complexity.

Hybrid/virtual partitioning is a JCAC scheme, which manages to combine the advantages of CS and CP and strikes a balance between unrestricted sharing in CS and unrestricted isolation in CP. More specifically, hybrid/virtual partitioning scheme behaves like unrestricted sharing when the overall traffic is light and like complete isolation when the overall traffic is heavy. Hence, the best characteristics of CS and CP under different loadings are combined [152]. The general structure of a partition in hybrid/virtual partitioning is explicitly explained in [32]. More specifically, each partition has two main parts namely the "commonly shared" and the "reserved" area. Each partition is allowed to accept "external" service calls (i.e. calls which were initially aimed to be served by other partitions). However, this has to be performed in a controlled manner in order to prevent the flooding of the partitions with external calls. While an "external" service call can be placed only at the commonly shared area, a "native" service call (i.e. a call which is mapped to be served by the specific partition) can be placed in any of the defined areas. By considering a dynamically changing reservation factor, the "commonly shared" area of the partition can be obtained by subtracting the "reserved" area capacity from the total capacity of the partition. As a result, as the reservation factor increases, the "commonly shared" area decreases thus accepting fewer "external" service calls and vice versa. It has to be noted that an accompanied preemption scheme is needed in order the robustness of the overall hybrid/virtual partitioning scheme to be supported. Hence, the main drawback is that the preemption scheme may lead to lower utilization.

Towards confronting this drawback, many supplementary algorithmic proposals have been made. Spillover-partitioning algorithms are proposed in [153], where utilization is improved by sharing certain partitions among different service calls. QoS degradation [121] [154] is also a well-known technique and is used in situations of network congestion. For example, when the network becomes congested, the amount of bandwidth allocated to some of the ongoing calls (also called degradable calls) is revoked to accommodate more incoming calls so that call dropping/blocking probabilities can be maintained at the target level without affecting resource utilization maximization targets. In case of light traffic load, some revenue maximization algorithms have been proposed such as those in [148]. In order to increase the resources utilization, some calls (also called upgradable) can be allocated with more resources. A typical example is web browsing or file downloading. In this scenario, the mobile user can enjoy better QoS (by first giving his consent for being excessively charged) and at the same time the operator can increase its revenues.

2.4.3 Advanced Context-Aware Approaches

As previously described, hybrid/virtual partitioning family of JCAC techniques seems be able to adequately and simultaneously satisfy most of the JCAC-related objectives in CAMoWiN area. However, research community envisions even more challenges regarding the future networking continuum, which need to be addressed. For example, JCAC-related architectural innovations proposed by various IEEE standards like P1900.4 [112] and 802.21 [111] stress the need for dealing with distributed radio resource usage optimization issues from an overall system perspective. Radio resource management in LTE-Advanced networks including related emerging challenges in HetNets [30], machine-to-machine communications (M2M) [155], device-to-device communications (D2D) [156] and cooperative communications [157] are also fields of research that are continuously gaining ground. Finally, novel JCAC principles have to be stressed for mobile cloud computing (MCC) environments as the idea of integrating cloud computing into heterogeneous mobile and wireless networking is promising, too [24] [158].

Regarding the afore-mentioned JCAC-related architectural innovations, the main objective is to define an appropriate system architecture and protocols, which will facilitate the optimization of radio resource usage by exploiting context information exchanged between network and mobile terminals, regardless of their support for multiple simultaneous links and dynamic spectrum access. The "Distributed Radio Resource Usage Optimization" use case introduced in [112], contains many context-aware JCAC-related building blocks, while reconfiguration and self-management features play a critical role, too.

Regarding LTE-Advanced innovations, the main breakthrough lies in the fact that the case of mobile terminals (MTs) being directly (i.e. via one hop) connected to base stations (BSs) of heterogeneous RATs in order to acquire access to services is not a panacea. In fact, many wireless/wired nodes can be relays of information, while small base stations (e.g. femtocells) can operate as relays, too [157]. Moreover, the concepts of M2M and D2D communications introduce the idea of MTs communicating directly with each other over M2M/D2D links, while remaining control under BSs. Due to this potential, location-aware and geo-referenced services can be developed and thus novel JCAC design has to be adopted.

In mobile cloud computing/networking, the main novelty feature, which has to be stressed is that a JCAC framework has to simultaneously take into account both: a) wireless/radio access resources pool and b) computing resources pool for data processing/storage aiming at flexible virtualized infrastructure sharing solutions. That is, there is no sense in allocating only networking resources to MTs, because there may not be corresponding sufficient computing resources to support the ongoing calls/services. Finally, joint design and optimization of access and backhaul networks is needed and hence JCAC modules have to be accordingly enhanced.

2.5 Ongoing Research on Networking and Computing Environments' Integration

As long as the integration of computing and networking environments is continuously boosted with the assistance of innovative ICT architectures and technologies, implications of Cloud Computing (CC) paradigm, which are applicable in mobile and wireless networking area are increasingly gaining ground [23]. Indeed, Mobile Cloud Computing (MCC) is an emerging research area and is introduced as the integration of CC into the mobile environment in order CA mobile applications to be supported [159]. The ultimate vision is to fulfill the dream of providing "information at everyone's fingertips anywhere at anytime" and as computation capabilities of mobile terminals (MTs) will always be a compromise, MCC aims at efficiently using CC techniques for data storage and processing on MTs, thereby reducing their limitations [127] [160]. Other major MTs-related technical restrictions are short battery lifetime, varying wireless channel conditions and high network latency, all of them hindering the remote display functionality of cloud applications on mobile devices [161]. More specifically, cloud-based mobile applications move data processing and storage away from mobile devices (MDs) to powerful and centralized platforms located in CC infrastructures. These centralized applications are then accessed over the wireless connection based on a thin native client or web browser on the MDs. From a general ICT perspective, MCC technology helps at: a) improving data storage capacity and processing power, b) boosting dynamic and context-aware service provisioning, systems' scalability, multi-tenancy and ease of integration c) improving reliability, security and privacy, d) offering more opportunities for new business models and business logic issues enhancement, and e) converging with transformative Future Internet (FI) technologies such as the Internet of Things (IoT) and the Internet of Services/Contents.

State-of-the-art MCC research proposes several architectural innovations. In [162], augmented execution of mobile cloud applications is proposed and thus CloneCloud proposes a method in which a replica of the application and data is kept in the cloud, and the computation is offloaded to the clone in the cloud. MAUI project [163] proposes dynamic partitioning of mobile cloud applications, where applications are partitioned and fine-grain offload of the code is sent to the cloud. Profiling information of the application, network connectivity measurements, bandwidth and latency estimations are used as input parameters to optimize and decide which piece and when it should be offloaded. In [127], a proxy-based mobile cloud method comprised by cloudlets is proposed, where MDs can use a nearby local server as a proxy between itself and the distant cloud. When MDs do not want to offload to the cloud, they can find a nearby cloudlet and thus mobile users may meet the demand for real-time interactive response by low-latency, one-hop, high-bandwidth wireless access to the cloudlet. Other frameworks dealing with MCC challenges such as seamless wireless

connectivity, context-aware services provisioning and security have also been proposed in [104], [23] and [164] correspondingly.

Apart from MCC, Mobile Cloud Networking (MCN) concepts are also emerging in the future converged networking and computing ICT continuum [165]. The basic assumption of MCN is the existence of Radio Access Networks (RANs), macro data centers (e.g. public cloud), micro data centers (e.g. community cloud), and cloudlets (private cloud). Macro data centers are standard large-scale computing server farms deployed and operated at strategically selected locations. Micro data centers are medium to small-scale deployments of server clusters across a certain geographic area, for instance covering a city or a certain rural area and as part of a mobile network infrastructure. It could also be owned and operated by business/institutional communities of specific interest (e.g. e-government, e-health, e-logistics, e-learning community, etc). A mobile network operator for instance may operate several RANs, mobile core networks, as well as data centers and thus enjoy full control of all technology domains. Another, more advanced example could be a company (could be a mobile network operator, a data centre provider, or any other enterprise) that acts as end-to-end MCN provider without owning and operating any physical infrastructure. This company would sign wholesale agreements with physical infrastructure owners, for instance mobile network carriers in those geographic areas it wishes to provide access to MCN services. The same would be the case for contracting data center operators in strategic locations and in order to complete a full MCN offering (RAN, Mobile Core, Data Centre). Conclusively, some of the main challenges that MCN research field should address are: a) QoS/QoE provisioning of mobile cloud resources, b) generally extend the concept of CC beyond data centres towards the mobile end user, c) provide integrated CA services by simultaneously managing mobile and fixed networking resources as well as computing and storage ones, d) enable novel business models, e) design flexible MCN architectures comprising of 4G and beyond HetNet topologies, various backhaul alternatives and flexible virtualized infrastructure-as-a-service (IaaS) solutions.

Finally, according to flexible MCN architectures challenge referred above, joint design and optimization of 4G Hetnet radio environment, heterogeneous backhaul (i.e. wired and wireless alternatives) and centralized processing should be addressed [166]. This solution will optimize the RAN system throughput and provide cloud-based mobile services instantly and efficiently in cost, energy, complexity and latency wherever and whenever the demand arises. The introduction of the Radio Access Network as a Service (RANaaS) concept has the potential to open the RAN/backhaul market for new players, like vendors and providers of cloud infrastructure platforms. RANaaS also provides the technological foundation for shorter and more efficient product development cycles due to the shift from dedicated equipment to software-based functions operated on cloud infrastructures.

As shown in figure 1.1, the PhD thesis consists of three main architectural pillars (aiming to deal with the aforementioned challenges from a context aware resource management perspective), namely: a) 4G HetNet environment in which radio access resources are efficiently managed, b) mobile and fixed networking systems' convergence points in which heterogeneous backhaul alternatives and integrated services QoS provisioning concepts are studied, and c) hybrid cloud infrastructure environment in which flexible virtualized IaaS sharing solutions are investigated. In the next three chapters, we focus on context aware resource management problem for the three architectural pillars correspondingly.

CHAPTER 3

CONTEXT AWARE RESOURCE MANAGEMENT IN A 4G HETNET ENVIRONMENT

In this chapter, context aware resource management aspects being applicable in a 4G HetNet environment are studied. According to figure 1.1 and regarding the three main architectural pillars of the current thesis, the scope of the research being undertaken in this chapter refers to the left hand side of the figure (i.e. 4G HetNet environment). More specifically, the term "4G HetNet environment" is initially defined accompanied by research challenges related with emerging small cell networks' technology. In section 3.2, the novelty features of the proposed femto-relay (FR) concept are introduced, while the assumed research problem is formulated. In sections 3.3 and 3.4, two proposed context aware frameworks for the femtocells' efficient integration in existing IP and cellular infrastructures are presented accompanied by related performance evaluation results. Section 3.5 provides further femto-relay concept's performance evaluation results regarding 4G HetNet system's area spectral efficiency enhancements, while section 3.6 concludes the chapter.

3.1 The 4G HetNet Environment

The Future Internet (FI) vision is expected to extend the "Always Best Connected" (ABC) notion and include use cases according to which the FI end user will attain any service on a single intelligent device (e.g. smartphone, tablet, etc) using any available network within a heterogeneous network environment consisting of open, cognitive and collaborative wireless and wireline networks. Since the early 2000s, when various wireless access network technologies were standardized (e.g. 3G UMTS, WLANs, WPANs, etc), it had become clear that the borders among the different RATs were quickly disappearing. At that time, the vision was that existing wireless network technologies could comprise a 4G all-IP overlay architecture, in which the coverage of different technology cells overlap with each other [117]. For example, a lot of research efforts have been undertaken for cellular/WLAN, WLAN/WiMAX, DVB/WLAN, WPAN/WLAN and other related interworking scenarios. During the last years, 3GPP Long Term Evolution (LTE) and LTE-Advanced (LTE-A) standardization groups are working towards stressing the 4G Heterogeneous Network (HetNet) research challenges [30].

4G HetNet deployment paradigm is an emerging research trend considered as the most significant aspect for meeting the continuous increase in mobile data traffic, mainly by proposing spectral efficiency enhancements and cell capacity gains from an overall system perspective. In order to improve the 4G network's efficiency, there is a mentality shift from traditional static cellular deployments towards multi-tier architectures, where small cells (i.e.

microcells, picocells, femtocells) are overlaid by larger ones [167]. In other words, 4G HetNets try to dynamically leverage the existing cellular network topology in order to increase the proximity between the access network and the end users. HetNet topologies include distributed antennas systems connected through fiber or microwave technologies with macro BSs, picocells which are miniaturized BSs, relays which route data through multi-hops and femtocells [30]. The latter are access points (i.e. home BSs), which can be installed by users in a plug-and-play fashion, while the mobile data traffic is routed via the wired xDSL backhaul. Femtocells have the potential to provide a win-win situation for users and cellular operators. From the users' perspective, the increased indoor coverage can provide services of high quality of experience (QoE) within the boundaries of their local network. On the other hand, the adoption of femtocell technology by users leads to a considerable decrease in capital and operational expenditures (CAPEX and OPEX) for cellular operators, while better traffic load balancing, overall capacity gains and area spectral efficiency can be achieved, too [168].

To achieve these gains, detrimental factors, which are introduced by the multi-tier cellular architecture, must be tackled. A number of recent studies indicate the interference arising between the tiers of the network (cross-tier interference), as well as the one among the same tiers (co-tier interference), as the main challenge threatening the efficient integration of femtocells in the cellular network [169] [170] [171]. Another major challenge that arises due to the femtocell architecture is the limited and QoS-unreliable wired backhaul. More specifically, the latter is inherently not equipped to provide delay resiliency and thus operators need to plan HetNets' backhaul carefully in order an optimal mixture of both wireless and wired backhaul alternatives to be realized [30] [168].

3.1.1 Challenges in femto/small-cell networks

In a femto/small cell setup, challenges that were not present in traditional cellular deployments result in degradation of network's performance. Figure 3.1 depicts the main challenges in a network consisting of a macrocell and a number of femtocells, which use the same spectral resources. Each femtocell consists of one home BS (HBS) and an arbitrary number of mobile terminals/user equipments (MTs/UEs). In this setup, resources are allocated with a frequency reuse factor of one, as there is no consideration of a separate subcarrier allocation between tiers. Having this in mind, the dense deployment of HBSs and MTs/UEs will result in interference between two-tiers (e.g. cross-tier macro-femto interference) or among the same tiers (co-tier femto-to-femto interference). Another factor, often blamed for being a serious drawback for femtocell deployments, is the backhaul that is based on xDSL technology. As femtocells were developed bearing in mind the positive effect of offloading macrocell BSs, femtocell traffic is routed through the Internet. This fact may

result in unreliable QoS/QoE for the MTs/UEs due to the delay that is introduced as data traverse a network, where load fluctuations are unpredictable and beyond cellular operator's control. With the aid of the descriptive illustration in figure 3.1, the challenges mentioned below are further discussed.

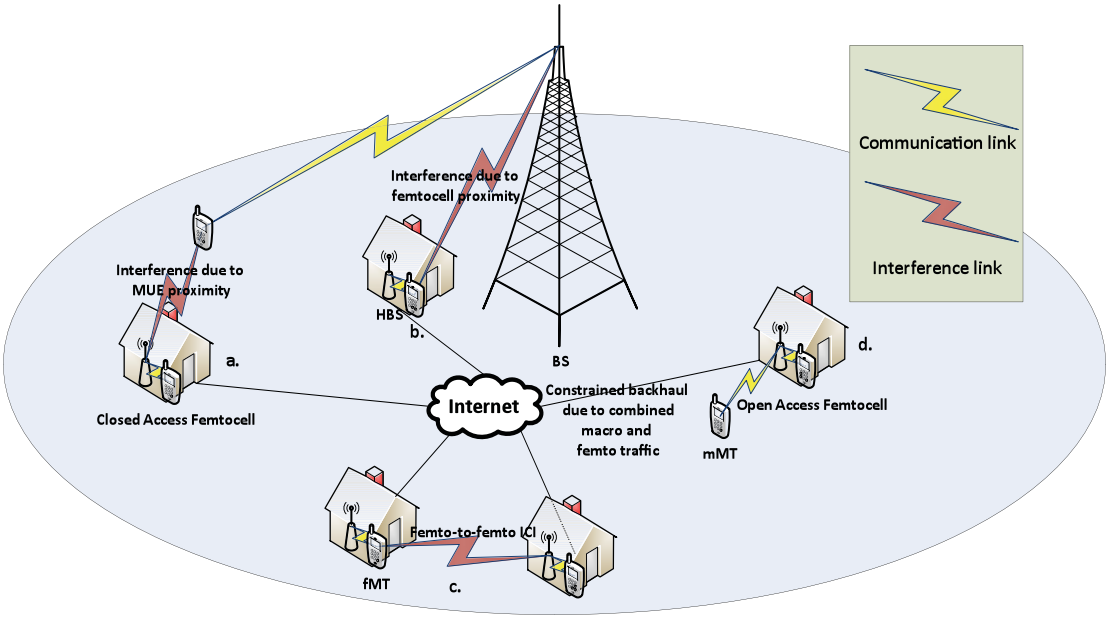

Figure 3.1: Overview of main challenges in a femto/small-cell network

Generally, the motivation behind the deployment of multi-tier networks is to bring access links closer to MTs/UEs, paving the way for a boost in spectral efficiency. Consequently, highly dense deployments are expected to be a common scenario for future cellular networks. Moreover, spectrum scarcity is pushing for non-orthogonal frequency allocation among tiers. These facts increase the probability for communication links, utilizing the same spectral resources to take place in proximity areas, thus experiencing severe mutual interference. Regarding cross-tier interference, the shortcoming comes from the overlapping of macrocell and femtocell coverage areas. In figure 3.1(a), we observe a closed access femtocell that has a macrocell MT (i.e. denoted as mMT communicating with macrocell BS) inside its coverage area. In a closed access femtocell, only a predefined set of users (i.e. closed subscriber group, CSG) can connect to the HBS. That is, only femto MTs (fMTs) can be connected, while mMTs cannot. In this case, HBS experiences mMT's signal as interference in the uplink and vice versa in the downlink. Power control [170] and beamforming [157] employed in both HBS and mMT could lead in improved performance in this case or better communication links provisioning to mMTs could protect them from interference [171].

Similar interference problems occur in figure 3.1(b)'s setting, where a femtocell access point (FAP) is near a macrocell BS. More specifically, the macro BS can incur interference to its nearby fMTs in the downlink degrading thus femtocell users' QoE. In this case, lowering the

transmission power of the macro BS can affect the QoS of the mMTs in the cell, so possible improvements could derive from power control in the uplink of femtocell MTs (fMTs) or by cancelling BS's interfering signal [27] [172].

Co-tier interference case, depicted in figure 3.1(c), arises when two femtocells are in overlapping areas. Beamforming techniques can be used to pattern HBS's radiation in such a way so as to a avoid "leakages" outside the femtocell area. Another technique is to employ a power control mechanism at HBS in order to dynamically adapt the range of its coverage and subsequently alleviate interference problems [168].

When open access policy is deployed (i.e. all fMTs and mMTs can obtain access to any HBS), femtocells avoid detrimental cross-tier interference by nearby mMTs. On the other hand, combined macro and femto traffic (i.e. from mMTs and fMTs correspondigly) puts an extra burden on the wired xDSL backhaul as shown in figure 3.1(d). As users of the local network may be connected using various interfaces such as LTE, Ethernet, Wi-Fi, WPAN, etc, backhaul capacity may not be enough to guarantee QoS levels promised in 4G networks. This limitation is also evident in hybrid (i.e. combination of closed and open access policies) and closed access femtocells, when the combined traffic of users served by wired backhaul approximates the latter's limited capacity [173]. Furthermore, QoS provisioning for delay-sensitive services can also become a serious concern due to the lack of wired backhaul network neutrality, even though this problem can be partially addressed in cases, where the wired backhaul provider is in tight strategic relationship with the cellular operator [168].

In section 3.2, a 4G HetNet architectural innovation concept called femto-relay is proposed, which addresses the above-mentioned challenges and can be easily integrated in HetNets' mobile broadband system design.

3.1.2 Problem Statement

Femtocells, while they are able to enhance the coverage and capacity of a cellular network with minimum cost by utilizing existing IP infrastructures, their unstructured placement may also cause major interference-related problems [31]. However, we argue that if the femtocell operation is assisted by relays, then the signal transmissions to/from the macrocell can be moved closer to the MT. As a result, transmitting with better channel conditions leads to transmission power reduction in both uplink and downlink cases and so inter-cell interference (ICI) between the macrocell and the femtocells can be reduced. Consequently, femtocell technology, if fully exploited, can lead to improvements in: a) interference reduction and b) data rate means, not only for the femto-users (fMTs/fUEs) but also nearby macro-users (mMTs/mUEs).

Conclusively, there is a definite need for cooperation between macro and femto base stations. In most cases, femtocells should be operated in open or hybrid access mode as this can

significantly improve the overall performance of the 4G HetNet. However, to the best of our knowledge, none of the works in the field considers a wireless relay link between the macro BS and the Femtocell Access Points (FAPs) and, even more, a femtocell equipped with a relay module. Thus, in the following we study this concept (section 3.2) and at the same time we provide the framework for the efficient cooperation of macro BS and femto-relays dealing with ICI problem for both uplink and downlink cases (sections 3.3 and 3.4 correspondingly). Finally, femto-relay concept's area spectral efficiency enhancements are shown for both unplanned and semi-planned femtocell deployment use cases.

3.2 The Femto-Relay Concept

The need for bringing access links closer to MTs and the mediocre indoor coverage are among the main motivations for 4G HetNets' technology development. Femtocells were mainly deployed for two reasons: a) better load balancing due to the use of xDSL for data routing and b) increased indoor coverage. On the other hand, wireless relaying [174] has been introduced in 3GPP LTE-Advanced and IEEE 802.16j for WiMAX specifications as a means to maintain a uniform and fair QoS for all users in a wireless network thanks to multi-hop transmissions. Furthermore, wireless relays deal with the detrimental effects of the wireless channel (i.e. pathloss, fading and shadowing). Apart from outdoor deployments, relays have also been suggested for indoor environments as expensive and time-consuming wired networking is avoided. Another benefit is that data is routed through the cellular operator's infrastructure easing thus QoS control.

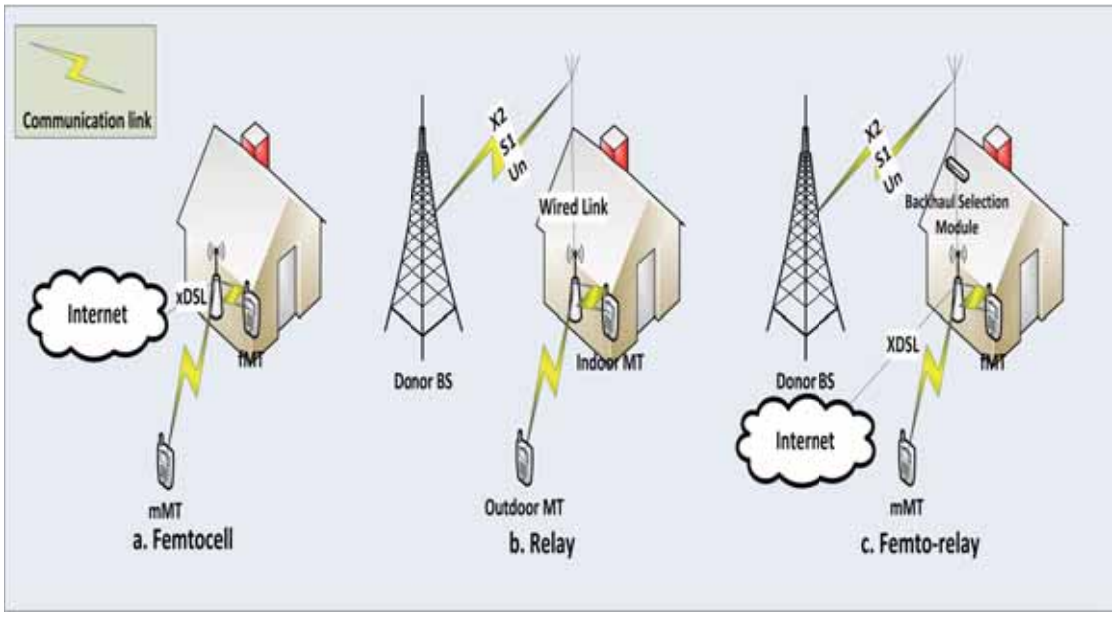

Figure 3.2: Comparison between (a) femtocells, (b) relays and (c) femto-relays

From this brief description, femtocells and wireless relays seem to have diverse characteristics due to their different originations, but still, they target similar goals from a 4G HetNet's perspective. One can easily infer that it is natural for operators and service providers to perform a tradeoff analysis in order to decide which technology better suits their needs and to follow a specific deployment strategy. On the other hand, a hybrid solution could combine the best of both worlds aiming to provide an optimal solution for future cellular networks. We name this solution as femto-relay concept and in figure 3.2, a comparison between femtocells and relays in indoor environments is depicted.

At the left hand side of figure 3.2, the classical indoor femtocell deployment is depicted. A femtocell access point/home base station (FAP/HBS) is installed indoors. Several MTs are communicating with the femtocell both from the closed subscriber group (CSG) of users (fMTs/fUEs) and other non-CSG users (mMTs/mUEs), who are in the vicinity of the FAP. The combined mobile data traffic both from fMTs and mMTs is routed via the wired xDSL backhaul. In case 3.2(b), the wireless relay concept is depicted. All MTs/UEs (both indoor and outdoor), which experience bad wireless channel conditions regarding their direct communication link with the macro BS, can communicate via a two-hop transmission with the latter (i.e. MT → relay → macro BS). Hence, better overall communication conditions can be achieved in cases when the MT experiences good channel quality with the FAP (i.e. first hop) and the wireless relay module of FAP experiences good channel quality with the macro BS (i.e. second hop).

The main differentiation point of femto-relays (see figure 3.2(c) case) is that their backhaul consists of a combination of wired and wireless communication links. Moreover, the supported interfaces are the same with the ones considered for femtocells and relays, so no further deviations from currently standardized equipment are required. Another basic element is the backhaul selection module (BSM) that operates based on the goals that have been set by the femto-relay's owner or by the interested organization/company in cases where femto-relays are installed by IT professionals. It has to be noted that the proposed femto-relay concept substantially differs from all other related works found in the international literature (see section 3.2.2).

3.2.1 Overview of femto-relay functionalities

Here, an overview of the main femto-relay functionalities is given. Figure 3.3 presents in detail the gains that femto-relays introduce to the network due to their double functionality as both femtocells and relays.

The ways that a femto-relay deployment can help in interference protection include both downlink and uplink scenarios (cf. figure 3.3). In downlink, an mMT that has a weak direct link with the BS and is located in the vicinity of an open-access femto-relay can have an

increased SINR if it selects to cooperate. Hence, the number of situations where two-tier interference could lead in outages can be reduced. In uplink, an mMT that transmits with full power to maintain its connection with the macro BS, can instead choose a two-hop transmission through an open-access femto-relay. This cooperation will result in power reduction by the mMT and incurred interference in nearby closed-access or non-selected femto-relays will be mitigated. The addition of the wireless backhaul alternative offers an additional degree of freedom in cooperation among femtocells and mMTs as the wired backhaul limitation can be alleviated. Therefore, the femtocells have the incentive of cooperation as backhaul capacity shortages are reduced and the reduced interference from and to the mMTs can increase the QoE of both fMTs and mMTs.

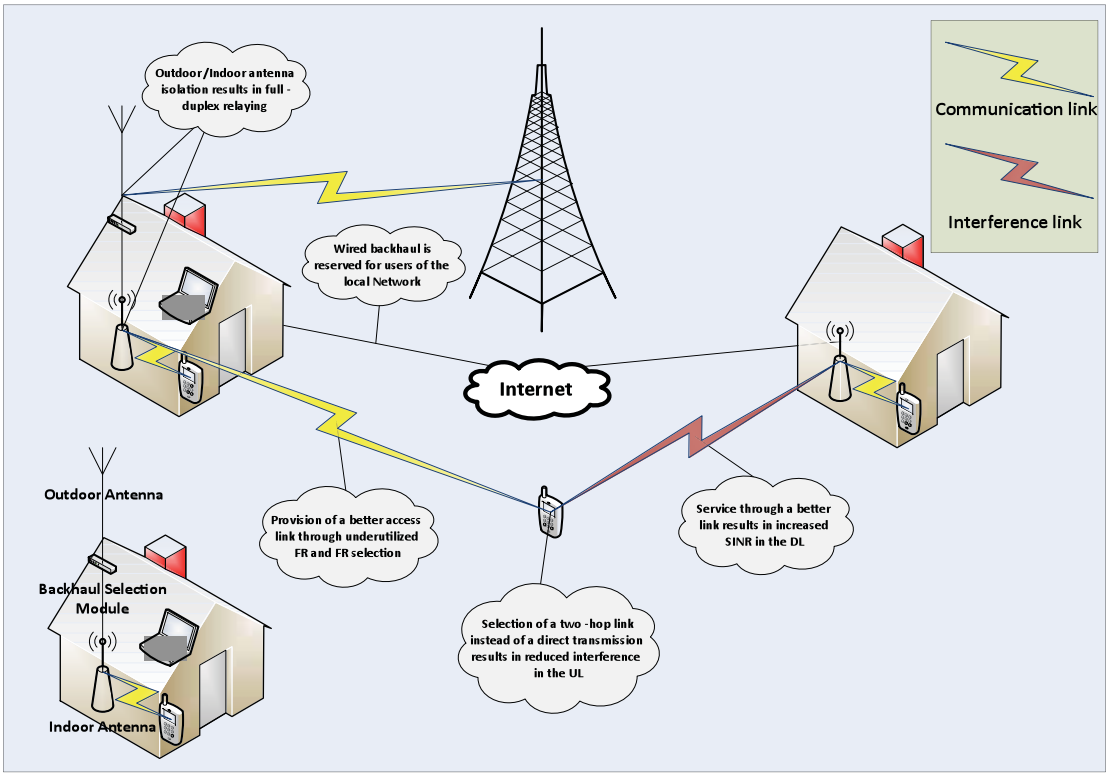

Figure 3.3: Overview of Femto-Relay (FR) Functionalities

The main novelty of femto-relaying is the two-fold backhaul alternatives, that consider routing of mobile data traffic either through xDSL connection and the Internet (i.e. wired backhaul) or directly to the donor BS (i.e. wireless backhaul) that offers the best SINR. In [172] and [175], the addition of a wireless backhaul to the currently deployed femtocells was proposed. The main differentiation was that in [172], the mMT data was routed only through the wireless backhaul thus providing the incentive for the installation of the wireless backhaul while satisfying privacy concerns as the private owned wired backhaul is avoided. On the other hand, in [175], routing decisions are triggered when a QoS threshold for the wired backhaul isn't satisfied leading to the use of the wireless backhaul by both fMTs and mMTs.

As a result, a multitude of challenging scenarios can be supported and interesting routing policies can be developed according to the combination of access policy employed at the available FAPs, too. For example, the administrator of an open or hybrid access femto-relay can parameterize BSM to exhibit a context-aware prioritization based on the type of services that the femto-relay is currently supporting. If wired backhaul experiences intolerable delays (e.g. in overload-state situations or when delay-sensitive services cannot be effectively delivered), then the wireless backhaul is selected, as the required QoS levels can be more easily maintained, because of the fact that the data "remains" within the cellular operator's network.

Regarding opportunistic femto-relay selection, as density of future cellular networks mainly represented through 4G HetNet paradigm is expected to rise with a continuously increasing rate, the probability that MTs will be (more often) located in the coverage areas of multiple femto-relays will be increased, too. This fact introduces many opportunities for fading mitigation. As shown in [176], selecting the best relay partner in terms of SNR can lead to increased diversity and as a result, outage probability can be reduced. An alternative would also be to consider the problem from an overall network system perspective (i.e. macrocell) and thus enable opportunistic femto-relay selection with the help of a bias factor for better load balancing. More specifically, through this bias factor, the mMT could select the least loaded femto-relay instead of the one providing the best SINR, if the QoS threshold is still satisfied in the access link.

Finally, the functionality of full-duplex in-band relaying is applicable for proposed femto-relay concept. In many relay deployments, relays are assumed to operate in half-duplex mode as inadequate antenna isolation is provided for concurrent transmission and reception [174]. In the proposed femto-relay concept, the separation of transmit and receive antennas is a basic characteristic. Moreover, in order to achieve spectral efficiency, only in-band transmissions for reception and transmission are considered. In figure 3.3, we see that the antenna supporting the wireless backhaul is installed on the roof of the femto-relay site, while the access antenna is located indoors. This outdoor-indoor separation can provide enough antenna isolation to guarantee a full-duplex relaying operation. For example in downlink, the outdoor antenna will receive the current frame from the donor BS, while the indoor antenna can transmit the previous frame. As a result, the half-duplex constraint of conventional wireless relays can be surpassed and thus increased spectral efficiency can be achieved by femto-relay concept adoption. Regarding the femto-relay's identification in the network, from the MTs' side it appears as a standard BS while the donor BS handles it as a MT. This type of operation is similar to type 1.b in-band relays considered for LTE-Advanced, which are equipped with isolated antennas for reception and transmission.

3.2.2 Related work overview

In the classic femtocell setup, a femtocell relays the traffic of the femto users (i.e. fMTs/fUEs) utilizing the wired backhaul link. In more advanced cases, they could adopt an open access policy and cooperate with users served by the macro BS (i.e. mMTs/mUEs). In [170], the two different access policies are compared concluding that in TDMA/OFDMA networks, the decision of whether a macro UE (mUE) will be accepted in a hybrid access femtocell, depends on user density. In [157], authors suggest the use of femtocells as relays in the uplink stream of a cellular network by serving both indoor and outdoor users via the wireline backhaul. The decoder is positioned at the mobile operator network and soft information is provided by the femtocell/relay to facilitate decoding. Authors in [177], suggest utilizing users, in the range of the femtocell, as relays of information to the mUEs and conclude that with this system architecture, the macrocell's bandwidth shortage can be surpassed. [178] elaborates on the relaying concept of [177], employing femtocells with the ability to broadcast relay service availability to the mUEs on behalf of the femtocell users (fUEs). Moreover, it is shown that the proposed architecture achieves better load balancing for the macrocell and improved throughput for the mUEs. Furthermore, in [179], mitigation schemes using femtocells with relaying capabilities are presented. Finally, the proposed scheme of [180] introduces femtocells that are able to decode the control channel of the macrocell BS. In this way, the femto BSs can adjust their transmissions, since they have information about the scheduled users in the macrocell avoiding inter-cell interference with the macrocell. Additionally, interference cancellation is implemented at the femtocells as an extra interference mitigation measure.

So far, no work in the international literature studies the concept of merging classical relaying and femtocell technologies aiming to reduce the burden of the wireline backhaul. Consequently, we introduce the concept of femtocells as relays that can communicate with a wireless backhaul with the macro BS, allocating as a result, all the available wireline backhaul to the femto UEs (fUEs).

3.2.3 Interference management in the uplink

In this section, the focus of the study is on the interference management issues regarding the uplink case from the MTs/UEs perspective. Hence, we identify the macro UEs as the main source of interference towards the femto BSs. In figure 3.4, a reference scenario is depicted, where two macro UEs are in the vicinity of four femtocells (FAPs/HBSs). Aiming to improve the SINR at the femto BSs, we propose to take advantage of femto BSs that currently do not serve any femto UE and to utilize them as relays for the macro UE. In this way, reduction in transmission power of the macro UE compared to the power required for direct transmission towards to the macro BS can be achieved. By decreasing transmission power, neighboring

femtocells, which were not available to cooperate, will experience less macro interference, too.

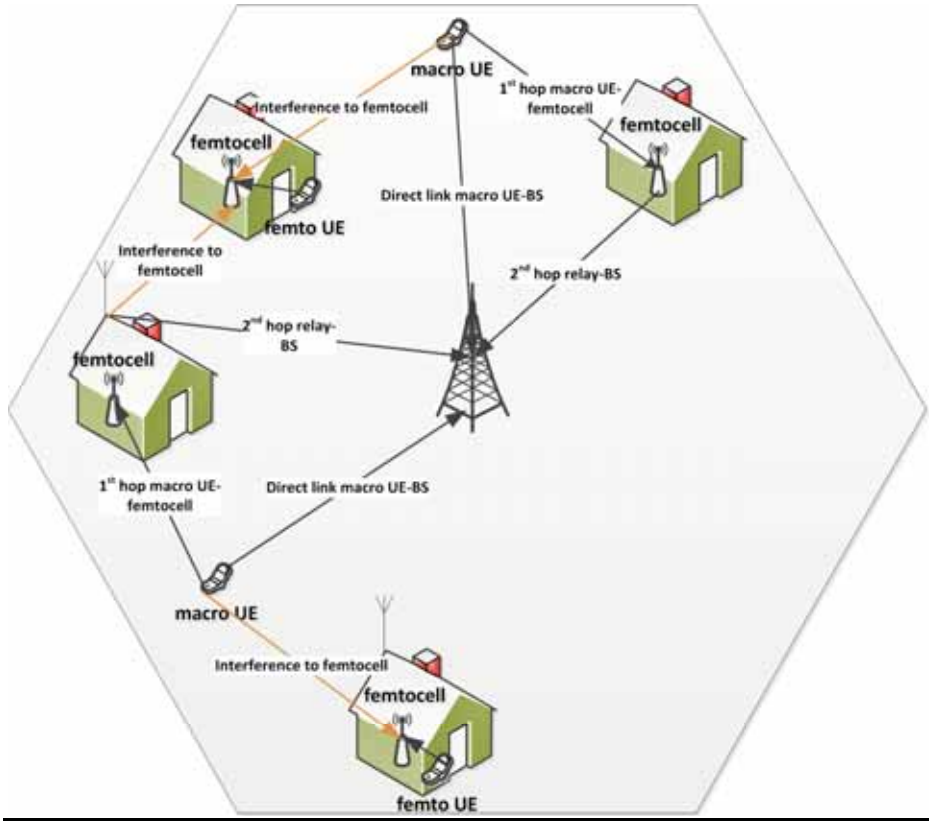

Figure 3.4: Interference management in the uplink case

Moreover, the deployment of outdoor donor antennas towards the macro BS is suggested, enhancing thus the femto BS with wireless relaying capabilities. The selection of outdoor antennas in the backhaul link increases the probability of having line-of-sight (LOS) conditions with the macro BS, resulting in reduced transmission power and less interference to nearby femtocells. In the case where no outdoor antenna is installed the indoor antenna could be used but with increased pathloss towards the macro BS. Furthermore, by choosing to route the macro UE traffic through the wireless backhaul, further protection and better usage of the wireline backhaul capacity of the femtocells is achieved. More specifically, indoor users connected with various interfaces (e.g. femto-cellular, WiFi, WPAN, Ethernet, etc) to an integrated services router won't have to share their backhaul capacity (e.g. xDSL capacity) with UEs, which don't belong in the CSG. In this case, the UE that is located in the femtocell's vicinity will receive a message denoting the latter's availability to cooperate. As the femtocell's location is fixed, it can acquire the CSI of the wireless backhaul link towards the macro BS with good accuracy. This information is included in the message sent to the UE together with the CSI of the femto BS-UE link. By receiving this information, the UE has to answer the following question: For the data rate required by its service, which route will

provide the largest power reduction? In order for the UE to decide its uplink route, a comparison between the direct and the two-hop links is performed. From the above description, one may see that the exploitation of the spatial diversity provided by the deployment of femtocells could lead to a two-hop transmission requiring reduced transmission power compared to a direct transmission towards the macro BS. As a result, having an additional path to choose from may effectively reduce the interference arising from the transmissions of the macro UE.

At this point, it needs to be outlined that the reader should combine information with chapter 4 of the current thesis in order to be able to understand femto-relay concept operation in conjunction with the various proposed backhaul protection and integrated QoS provisioning schemes presented therein.

3.2.4 Interference management in the downlink

In this section, the focus of the study is on the interference management issues regarding the downlink case from the MTs/UEs perspective. As shown in figure 3.5, a single macrocell in an urban environment is assumed, where the macro BS is located in the center. Within the area of the cell, a number of femtocells are also deployed. It is also assumed that some of the femtocells are able to communicate with the macro BS through a wireless relay link. To enhance the benefits of wireless connectivity, some femtocells may be equipped with outdoor donor antennas to support this relay link. The macrocell's users are moving in the cell in a random fashion and in random instances they may be within one or more femtocells' coverage area.

In the downlink, when the macro BS is transmitting and mUEs are receiving, the users served by femtocells (i.e. fUEs) may experience increased interference if the signal from the macro BS is strong. Furthermore, when the femtocells are serving their users, macro UEs in the vicinity of the femtocells are interfered in the reception of the macrocell BS's downlink transmissions. This setup is illustrated in figure 3.5, where the interference signals towards the femto UEs and macro UEs are highlighted with orange color, while communication links are highlighted with black color.

The proposed femto-relay concept aims at reducing ICI at the downlink of femtocells that are located within a macrocell's coverage area. Hence, context information about underutilized femtocells can be exploited in order neighboring mUEs to be served. However, as the femtocells' IP backhaul may be utilized by users connected through Ethernet/Wi-Fi/WPAN interfaces, it is suggested that the traffic destined to the nearby macro UE should be directly relayed through a wireless link, from the macro BS to the femtocell, as is the case of typical relaying. More specifically, the macro UE receives messages from the nearby femtocells denoting their availability to serve as relays. Estimating the SINR of all the available links,

including the direct link to the macro BS, the most suitable path for each macro UE can be opportunistically selected.

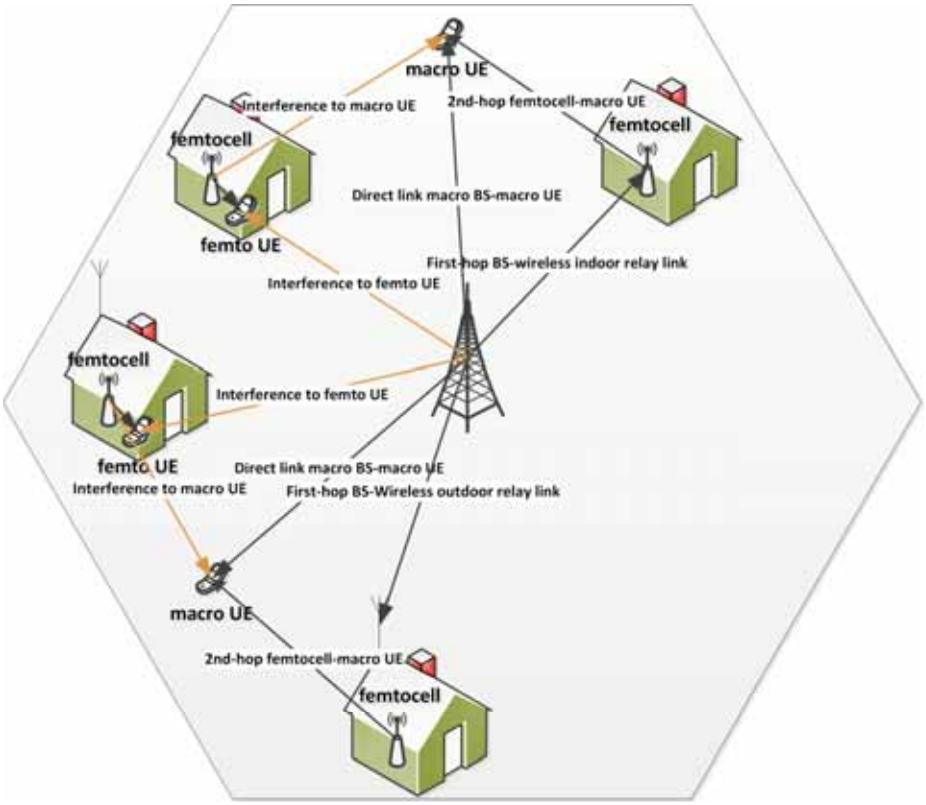

Figure 3.5: Interference management in the downlink case

In sections 3.3 and 3.4, two frameworks, namely CA-FEI and COF-FEI, for the efficient integration of femto-relay concept in the 4G HetNet environment are proposed. The former deals with interference management in the uplink case, while the latter deals with interference management in the downlink case.

3.3 Proposed CA-FEI Framework

In this section, a context-aware framework for the efficient integration of femtocells in IP and cellular infrastructures (CA-FEI) is introduced. CA-FEI framework deals with the interference management issues in the uplink case, as those are presented in 3.2.3. It should be noted that in this section only the femto-relay concept's related functionalities are studied, while CA-FEI framework includes some other context aware resource management modules for mobile and fixed networking systems' convergence, which are presented in chapter 4.

Prior to the description of CA-FEI framework, a typical topology of a femtocell deployment is provided, which helps the reader in clearly pointing out the proposed architectural modifications. As shown in figure 3.6, the main network elements of a typical femtocell deployment are the Femtocell Access Point (FAP) itself and a broadband IP router that allows

FAP to access the internet and consequently the Femtocell GateWay (FGW), which serves several FAPs. The communication between the FAP and the FGW is performed through the Iu-h interface while the Iu-Cs (Circuit switched) and Iu-Ps (Packet switched) interfaces are used for the communication between the FGW and the Mobile Core Network [181]. As it also becomes apparent from network topology depicted in figure 3.6, the router may also serve a number of wired LAN users as well as a number of wireless users through an attached Wi-Fi Access Point (note that any other wireless access interface may be available in the integrated router, e.g. WPAN-like interfaces). Furthermore, the FAP should be able to serve concurrently, according to the scenario being assumed [32], from four up to thirty-two UEs and offer them the same services and the same QoS as if a typical NodeB served them.

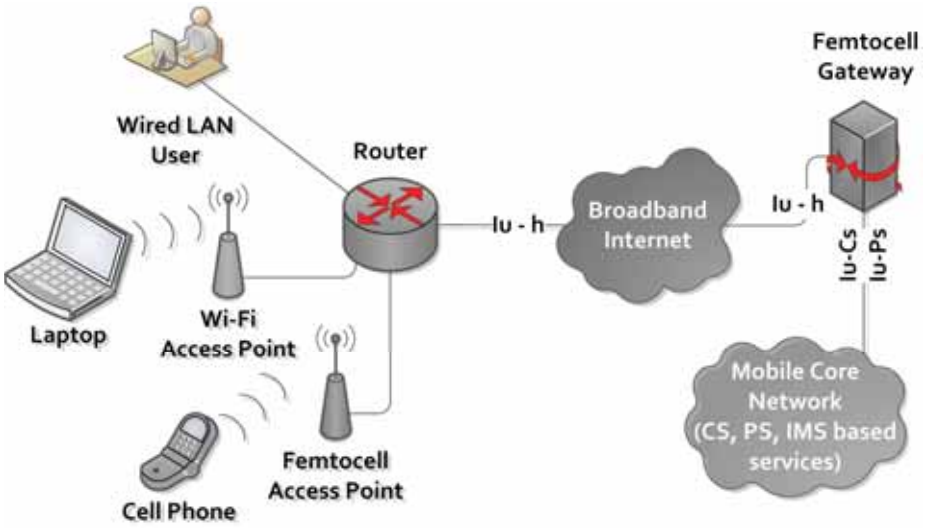

Figure 3.6: Typical topology of a femtocell deployment

The proposed CA-FEI scheme (see figure 3.7) is practical, as it does not require any modifications to the existing IP infrastructure. Actually, it can be realized as an add-on unit, placed on top of the existing networks, in the form of an integrated IP Router/Femtocell Access Point (FAP). In this section, we focus on the "relay module" depicted in figure 3.7 considering "context information acquisition module" (CIAM) and "RRM integration layer" as black-box functionalities, which will be presented in chapter 4. More specifically, the decode-and-forward (DF) relay module gives the ability to the assumed integrated router/FAP to act, when possible, as a relay for the macrocell users. This can be done in cases when some performance evaluation metrics employed in CIAM and RRM integration layer's modules satisfy specific thresholds and constraints allowing thus the relay module to be activated in order better communication link conditions to be realized.

Contrary to the classic setup (see figure 3.6), an additional functionality implemented in the proposed framework is DF relaying. Therefore, wireless backhaul links are used towards the macro BS by adding an outdoor donor antenna connected through cable link with the femto

BS. In this way, wireless relaying capabilities to the femtocells are integrated in order to protect the communication of the femtocell BS and the femto UEs using the wired backhaul link in terms of QoS. Furthermore, as the macro UE communication with the femto BS will be performed on small distance links, there is an increased possibility of achieving a reduced power transmission in this part of the macrocell. This fact can lead to better SINR in other femtocells in the vicinity of the macro UE, which were not available to cooperate at that time (cf. section 3.2.3). Moreover, as the second hop is performed via an outdoor antenna on the rooftop of the building, there is a good chance of having LOS condition with the macro BS. As a result, the required power in the second hop can also be reduced, while attaining the desired data rate in the end-to-end two-hop link.

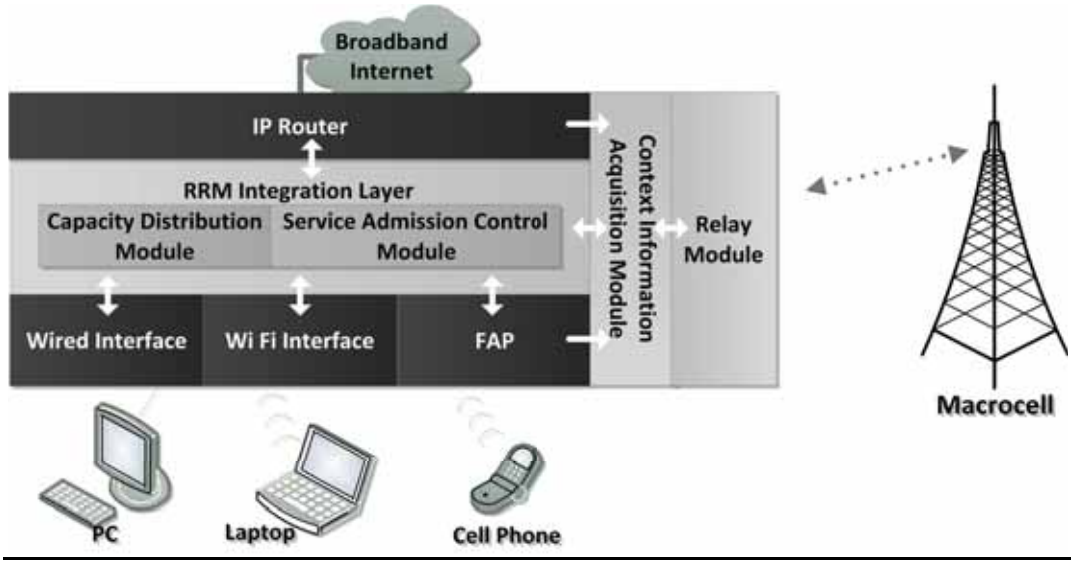

Figure 3.7: The proposed CA-FEI framework

3.3.1 Simulation environment setup

In our system, only pathloss and log-normal shadowing effects are assumed according to the suggested parameters in [182]. The frequency used for the communication is 2 GHz, the height of the macro BS antenna is 32 m, while that of the relay node (RN) is 10 m. For the indoor antenna, a height half to that of the RN at 5 m is considered and finally, the UE antenna's height is at 1.5 m. Also, each link exhibits a different shadowing standard deviation as shown in Table 3.1. Moreover, for the link between the UE and the indoor femto BS antenna, we consider that an additional penetration loss equal to 20 dB is added in the link budget calculations. Following the technical specifications of [182], the pathloss of the direct link between a macro UE and the macro BS is given by:

$$PL(R)_{\text{UE-BS}} = 131.1 + 42.8\log_{10}(R) \qquad (3.1)$$

where R is the distance in kilometers between the transmitter and the receiver. The shadowing standard deviation is equal to 10 dB. In equation (3.1), we take under consideration the non-

line-of-sight (NLOS) case, which in an urban setting of UEs, maintains, most of the time, NLOS connectivity with the macro BS.

Table 3.1: Simulation parameters for CA-FEI framework

Parameter	Value
Macrocell radius (m)	500
Femtocell radius (m)	50
Number of femtocells	0–100
Number of macro UEs	25
Modulation	Adaptive modulation (QPSK, 16 QAM)
Shadowing std UE-Macro BS (dB)	10
Shadowing std UE-Femto BS (dB)	10
Shadowing std femto BS-macro BS (dB)	6
UE transmit SNR (dBm)	23
Macro BS transmit SNR (dBm)	46
Femto BS transmit SNR (dBm)	23
Receiver diversity gain (dB)	3
Penetration loss (dB)	20
Results	Power reduction, macro UE data rate

For the two-hop transmission, in the link between the macro UE and the cooperating femto BS, NLOS conditions will be dominant, as the access antennas of the femtocells will be located inside buildings. As a result, the pathloss will be:

$$PL(R)_{\text{UE-BS}} = 128.1 + 37.6\log_{10}(R) \quad (3.2)$$

The shadowing standard deviation has a value of 10 dB. In the second hop, when an outdoor antenna is installed, the pathloss of the femto BS–macro BS link is calculated as follows:

$$PL(R)_{\text{fBS-BS}} = 100.7 + 23.5\log_{10}(R) \quad (3.3)$$

In this case, we assume line-of-sight (LOS) conditions as the outdoor antenna will be on the rooftop and directed towards the macro BS. The shadowing standard deviation is equal to

6 dB. On the contrary, when no outdoor antenna is used, the link is assumed to be NLOS and the pathloss is:

$$PL(R)_{\text{BS-BS}} = 131.1 + 42.8\log_{10}(R) \quad (3.4)$$

In this link, the shadowing standard deviation has a greater value, equal to 10 dB.

In order to study the efficiency of the proposed CA-FEI framework and more specifically the efficiency of the proposed DF relay module in the uplink case of a cellular network, a simulator in Matlab was developed considering [182] for the simulation parameters (cf. table 3.1). We consider a macrocell with radius 500 m operating in an urban environment with the macro BS located in the center and a varying number of femtocells that are available for cooperation. The macro BS is equipped with two receive antennas providing a diversity gain of 3 dB. The femtocells have a coverage radius of 50 m and are uniformly located in the macrocell. The reference receiver sensitivity values are calculated using the values in table 22.6 of [183], considering a bandwidth of 20 MHz and the following equation:

$$REFSENS = kTB + \text{NF} + \text{SINR} + \text{IM} - 3(\text{dBm}) \quad (3.5)$$

where *kTB* is the thermal noise level equal to –174 dBm/Hz, NF is the prescribed maximum noise figure for the receiver and IM is the implementation margin while –3 dB is the diversity gain. The modulation schemes range from QPSK 1/3 to 16QAM 4/5 and we use the corresponding SINR and IM values again from Table 22.6 of [183]. As mentioned above, the femtocells are assumed to have an outdoor donor antenna to support the wireless backhaul link towards the macro BS and a classic indoor coverage antenna. For simplicity, we assume NLOS conditions for UE-to-macro BS and UE-to-femto BS and LOS conditions for the femto BS-to-macro BS links. The UEs are also dropped uniformly in the macrocell area and are equipped with one transmit antenna. As a result, the maximum modulation order they can support is 16 QAM. Whenever a two-hop transmission is selected, a penetration loss of 20 dB is considered [182]. Also, overhead due to signaling when a macro UE is hand-offed from the macro BS to a femto BS and vice versa is considered negligible in the data rate results that are presented in the following paragraphs.

3.3.2 Performance evaluation results

To evaluate the performance of the proposed DF relay module, we simulate the above-mentioned topology to obtain numerical results regarding: a) power reduction in the transmission of the macro UEs and the backhaul antennas, and b) data rate achieved by them. For the first set of numerical results, we set a target data rate of 14.4 Mbps for the uplink as it is a reasonable value considering QoS fairness even for cell edge users. As a reference scheme, the classic femtocell setup is used, where only users of the CSG are served and there is no capability to cooperate with other UEs. Results of the simulation are depicted in

figure 3.8 for the increasing percentage of available femtocells. It should be noted that the values for power reduction are the differential gain of the proposed solution (i.e. adopting the femto-relay concept) compared to the classic femtocell architecture (i.e. considering FAP without femto-relaying functionalities).

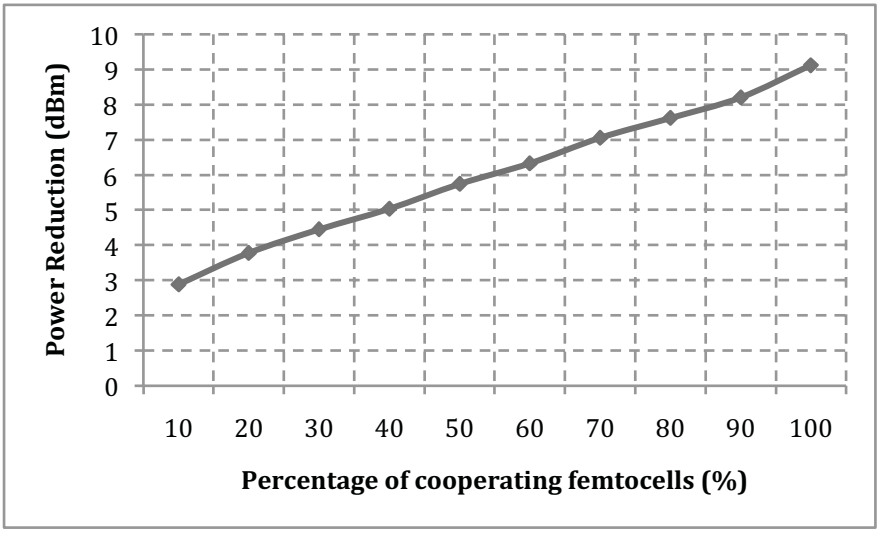

Figure 3.8: Power reduction for varying percentage of femtocell cooperation

One may observe that as the available femtocells increase, there is a greater chance to achieve a transmission with reduced power by the interfering macro UEs and the backhaul antennas positioned on the roof of the building. The decrease in dBm ranges from 2.9 to 9.07 dBm for the extreme case, where all the femtocells are available (cf. figure 3.4). For medium cooperation factors (i.e. 30–50%), we see that power reduction takes values between 4.5 and 6 dBm. As a result, to achieve the specified data rate, macro interference can be reduced whenever an additional path is available to select from (i.e. when a cooperating femtocell is near a macro UE). Moreover, the incurred interference to femtocells currently serving UEs of CSG will be accordingly reduced.

In the next simulation scenario, we assume the cases of 30% and 40% availability of cooperating femtocells and we examine the possibility of performing the second hop towards the macro BS through the wireline backhaul in case it can be allocated for macro UEs traffic. Figure 3.9 shows the results for varying percentage of wired transmissions by the cooperating femtocells for the link between the femto BS and the macro BS. From Figure 3.9, one may observe that for each case, if the wired backhaul is available, we completely avoid the interference in the second hop. As a result, we see an improvement over the case of only wireless backhaul, which for the case of 30% has a range between 4.4 and 6.15 dBm and for the case of 40% between 5.1 and 7.3 dBm.

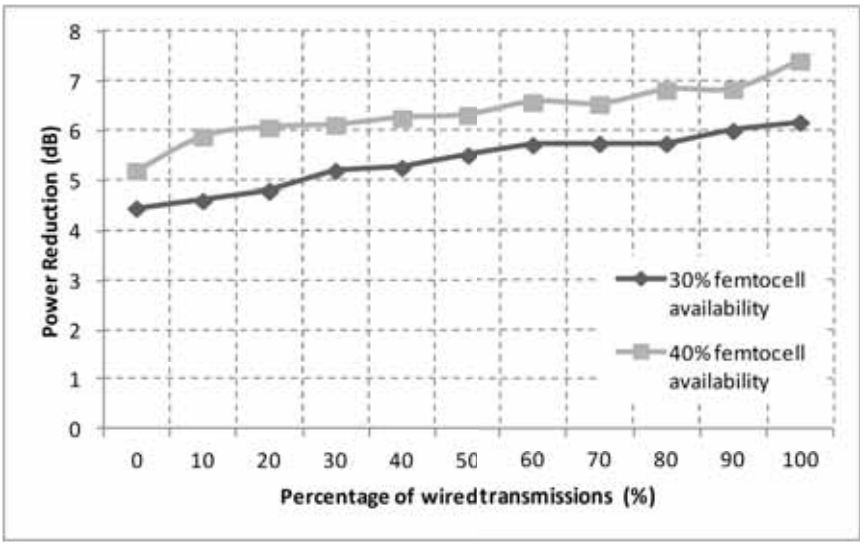

Figure 3.9: Power reduction for varying percentage of wired transmissions for the cases of 30% and 40% femtocell availability

In figure 3.10, we present the numerical results for the data rate improvement of macro UEs in order to examine whether CA-FEI framework can provide gains in this area, compared to the classic femtocell setup. In order to study the increase in the data rate of the macro UEs, the latter adopt a policy of selecting the best end-to-end path and transmitting with the best possible data rate. This comes in contrast to the power reduction policy of the previous simulation. To measure the end-to-end capacity, the half-duplex constraint is employed in the two-hop transmission and the capacity is given by [184]:

$$C_{\text{two-hop}} = \frac{1}{2}\min\left(C_{\text{first-hop}}, C_{\text{second-hop}}\right) \tag{3.6}$$

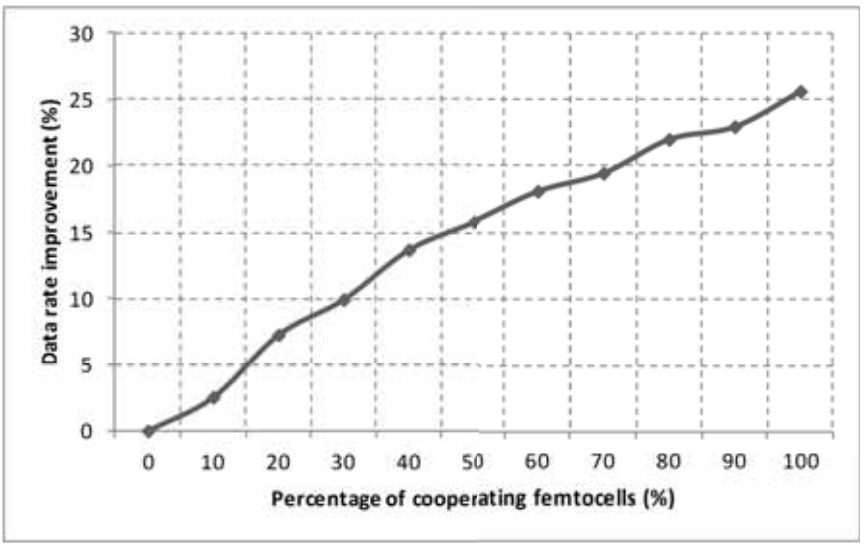

Figure 3.10: Macro UE data rate improvement for varying percentage of femtocell cooperation

One may see that when a macro UE receives an availability message from a femtocell with the best link, it can associate with it and relay its traffic through a two-hop link. As mentioned before, the two-hop route will be selected, when the direct link's data rate is below the two-hop's one. Hence, by exploiting spatial diversity, the macro UEs can improve their data rate by 26% in the extreme case of 100% available femtocells. For more realistic cases of 30–50% femtocell availability, the improvement can range between 10 and 16%. Again, it should be noted that these values are the relative improvement achieved by CA-FEI compared to the classic femtocell setup.

Finally, the case where an outdoor antenna for the wireless backhaul is not always available is studied. As one of the main benefits of femtocell deployment is their plug-and-play nature, studying the effectiveness of femtocell cooperation in cases where outdoor antennas are available in realistic percentages will give us an indication of the usefulness of the CA-FEI framework. More specifically, figure 3.11 presents the data rate performance of macro UEs in the cases of 30 and 40% availability of cooperating femtocells for a varying percentage of installed outdoor antennas.

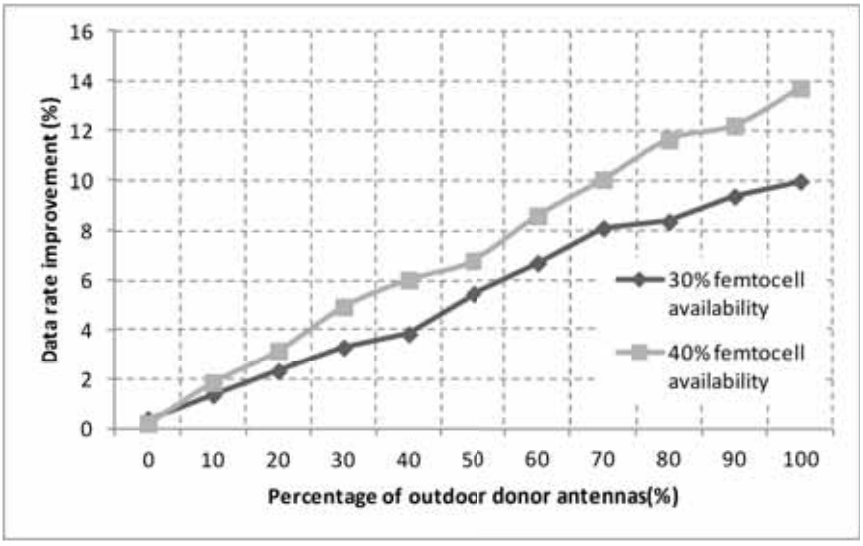

Figure 3.11: Macro UE data rate improvement for varying percentage of femtocell cooperation

One may see that macro UEs can benefit from femtocell cooperation in cases where the wireless backhaul is supported by outdoor antennas. When no outdoor antennas are available, data rate metric experiences almost no improvement due to the low transmit power of the femtocell and the penetration loss. On the other hand, when outdoor antennas are added to the network, we see significant improvement even for percentages below 50%. In the low percentage range, we observe that data rate increases between 1.3 and 5.5% for 30% femtocell availability, while for 40% femtocell availability, we see an increase of 1.8 to 6.7% compared to the case of only direct transmission in the uplink. Again, the above percentages

are relative improvement compared to the reference case of the classic femtocells without wireless backhaul.

3.4 Proposed COF-FEI Framework

In this section, a cooperation framework for the efficient integration of femtocells in IP and cellular infrastructures (COF-FEI) based on femto-relay concept is introduced. COF-FEI framework deals with the interference management issues in the downlink case, as those are presented in 3.2.4. It should be noted that COF-FEI consists of the same modules as those are depicted in figure 3.7 for the CA-FEI framework and in this section only the femto-relay concept's related functionalities are studied.

From the descriptions provided in 3.2.4 (cf. figure 3.5), one may easily conclude that the macro-femtocell interference is one of the main degrading factors in a typical 4G HetNet topology. COF-FEI's objective is two-fold: on the one hand, by reducing the power level of the macro BS's transmissions, the interference to femto UEs will be reduced. On the other hand, by providing the macro UEs with better links through open-access underutilized femtocells, their SINR will be enhanced, thus protecting them from interference by other femtocells that are currently serving their UEs.

More specifically, when a macro UE is in the vicinity of one or more femtocells that are available to cooperate, messages (i.e. context information) are transmitted from the latter to denote this event. Through this context information being exchanged, the macro UE is also informed about the conditions of the links between the macro BS and the femtocell as well as between the femtocell and macro UE. After receiving this information, the macro UE should decide on whether or not a two-hop transmission would result in a better SINR in comparison with a direct transmission from the macro BS. Finally, this decision is sent to the macro BS in order to relay the traffic destined to the macro UE through the cooperating proximate femtocell. Consequently, the latter activates its relay module (cf. figure 3.7) in order to be able to act according to the femto-relay concept being proposed in this thesis.

A Link Selection Algorithm (LSA) resides in the DF relay module of COF-FEI employing an opportunistic selection criterion [176], which is based on the SINR of each hop. LSA is performed at the macro UE side in order to define the most efficient link. Hence, the path with the best end-to-end transmission quality towards the macro UE is selected and thus better results in power reduction and data rate improvements can be achieved. Figure 3.12 presents a high-level description (i.e. pseudocode) of LSA.

It is evident that during system operation, LSA offers to the macro UE a transmission path that, in many cases, is better when compared to the direct link. In such cases, it is possible to lower the transmission power of the macro BS and at the same time to keep the SINR at the same, or better, level compared to the direct link. In other words, the 4G HetNet system can

keep the end-to-end capacity over a specific data rate threshold with lower transmission power. Consequently, that would result in less interference towards femto UEs. On the other hand, in a dense femtocell deployment, macro UEs having weak links towards the macro BS due to fading or cell-edge location could be effectively served through their neighboring FAPs.

FOR j=1:M macro UEs

FOR i=1:N available femtocells

$SINR_{best\ i} = \max\min\left(SINR_{BS-i}, SINR_{i-j}\right)$

IF $SINR_{best\ i} > SINR_{BS-j}$

Route the j^{th} macro UE's traffic through a two hop link

ELSE

Route the j^{th} macro UE's traffic through a one-hop link

END

END

END

Figure 3.12: Link Selection Algorithm

3.4.1 Simulation environment setup

The simulation environment setup for COF-FEI performance evaluation is similar with the one described in 3.3.1. The only difference is that COF-FEI deals with interference management issues in the downlink case (cf. section 3.2.4), while CA-FEI deals with interference management issues in the uplink case (cf. section 3.2.3). As a result, both frameworks can provide a complete context-aware resource management proposal being applicable in a 4G HetNet environment.

3.4.2 Performance evaluation results

In order to study COF-FEI's efficiency, a simulation setup was developed in MATLAB© and two simulation scenarios are considered. The details of the designated simulation's parameters are shown in table 3.1.

For the first simulation scenario, the ability of COF-FEI to reduce the transmission power of the macro BS is evaluated. As of the downlink case (i.e. the link from macro BS to UEs), power reduction can be achieved if the users experiencing bad communication links towards the macro BS are offered better two-hop links through nearby femtocells. Thus, we assume a cell consisting of a varying number of cooperating femtocells and a single service with target

data rate of 20 Mbps. We set a target data rate in order to quantify the power reduction achieved by COF-FEI for the corresponding SINR value [185].

A semi-planned deployment is also considered where apart from femtocell installation in houses, planned femtocell installation in office buildings, airports, train stations, malls, stadiums, city squares, campuses, etc could take place. Consequently, it is assumed that some FAPs are properly deployed by ICT specialists and equipped with outdoor antennas. The percentage of outdoor antennas is considered fixed at 50%, which is a reasonable value for planned femtocell deployments.

The results regarding power reduction vs. femtocell cooperation percentage are illustrated in figure 3.13. It can seen that as the percentage of available femtocells increases, more paths are offered to macro UEs, and so the required transmission power is reduced. The power reduction reaches a saturation point at 90% where any further increase of the number of available femtocells cannot offer a better transmission path to the macro UE.

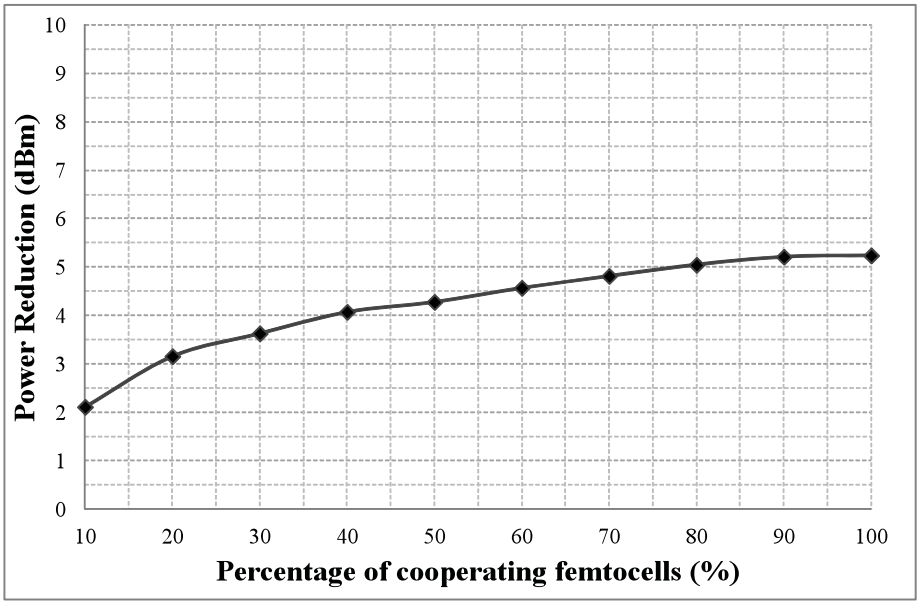

Figure 3.13: Power reduction vs. femtocell cooperation percentage

In the next simulation scenario, the percentage of the outdoor donor antennas is changed in order to study power reduction for the whole range of unplanned femtocell to fully planned femtocell deployments. We keep the percentage of femtocell cooperation fixed to 30% and 40%, which is a reasonable cooperation degree. The results are shown in figure 3.14. One may observe that the power is reduced in an almost linear fashion as the number of outdoor antennas increases. In this case, the percentage of the cell covered by the femtocells is stable and the ratio of macro UEs served by them does not vary, as is the case for the simulation involving the increasing percentage of cooperating femtocells. In the case of 40% cooperation, the corresponding curve has a higher slope indicating a greater range of

improvement compared to the 30% curve. For low percentages, the first-hop is the bottleneck of the end-to-end transmission in many cases, while for high percentages the NLOS second-hop is the main cause of degraded channel conditions.

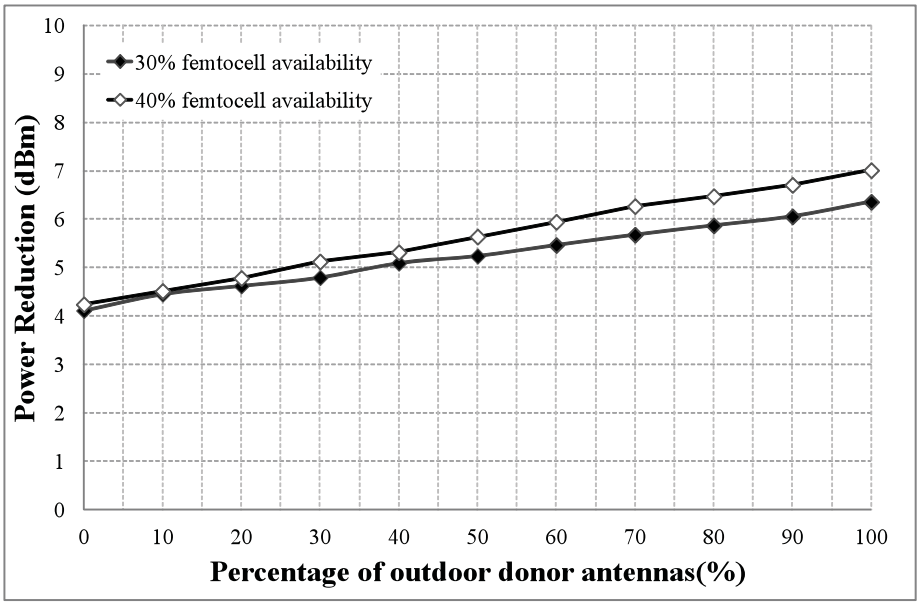

Figure 3.14: Power reduction vs. number of outdoor antennas

Furthermore, the effect of COF-FEI on the data rate of the macro UEs is examined. Assuming the same simulation scenarios as mentioned above, figures 3.15 and 3.16 illustrate the data rate improvement over the case of no femtocell cooperation when COF-FEI is employed, for varying percentage of cooperating femtocells. In these scenarios, no target data rate is set and the main link selection algorithm's (LSA) goal is to select the best path for each macro UE based on the achieved SINR, in order to maximize the end-to-end capacity. These simulations demonstrate a behavior similar to the corresponding comparisons for the power reduction. In figure 3.15, one may observe that the data rate improvement for percentages above 70% isn't as notable as for lower percentages. As shown previously, while more femtocells become available, more macro UEs tend to associate with them and for these values femtocell-macro UE association saturates. Finally, in figure 3.16, the relation of increasing the percentage of equipping the available femtocells with outdoor antennas is studied. Again, one may observe that for the 40% cooperation case, the slope of the curve is higher and the range of data rate improvement is greater.

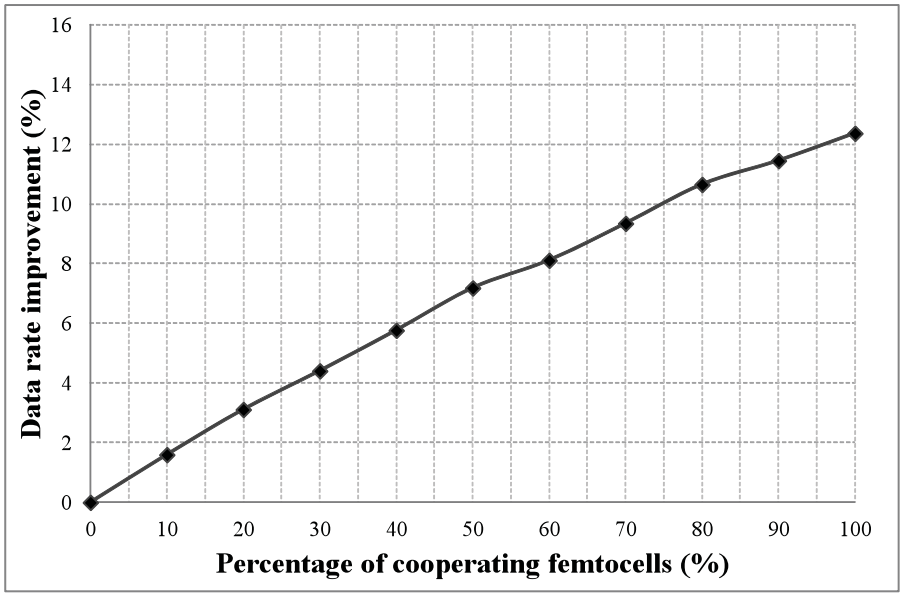

Figure 3.15: Data rate improvement vs. femtocell cooperation percentage (50% outdoor antennas)

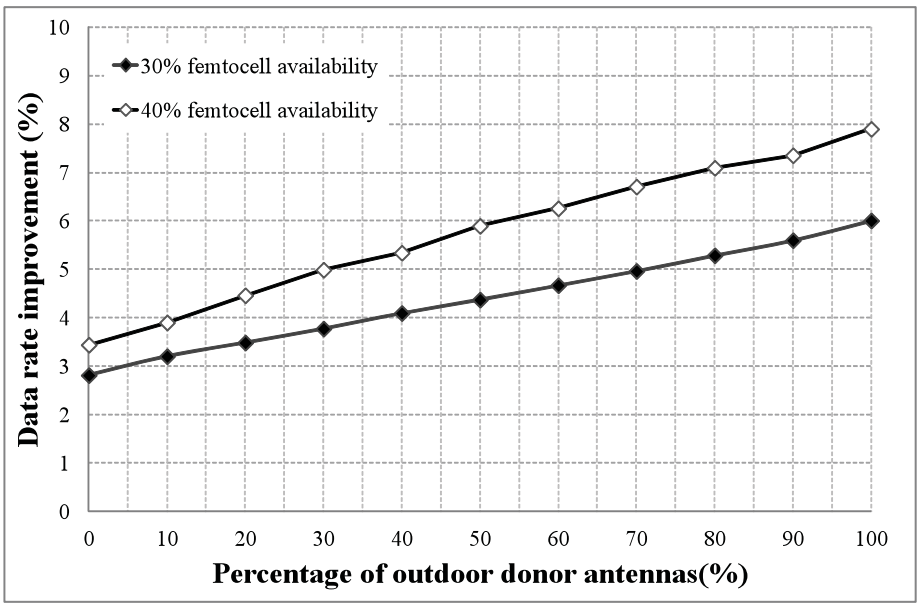

Figure 3.16: Data rate improvement vs number of outdoor antennas (30 and 40% femtocell cooperation)

3.5 Femto-Relaying Area Spectral Efficiency Enhancements

In this section, femto-relay concept's gains in terms of area spectral efficiency are studied. So, in order to further evaluate the anticipated technical impact of the proposed femto-relay concept, two femtocell deployment cases are assumed, namely unplanned and semi-planned, which both simulate a two-tier (i.e. macro-femto) cellular network architecture. In table 3.2, the main differential characteristics (e.g. positioning of tiers, percentage of open-access femtocells, UEs' density, etc) of the two deployment cases are sententiously presented. For

both cases, we study the relative improvement of area spectral efficiency for various UE categories, while an increasing femto-relays' penetration rate is assumed (i.e. the open-access femtocells are increasingly replaced by femto-relays, which have an additional wireless backhaul).

Table 3.2: Main differences between unplanned and semi-planned deployment cases

Unplanned deployment	**Semi-planned deployment**
One macrocell (i.e. unplanned area) and HBSs are randomly dispersed (in an unplanned manner)	One macrocell is divided in 2 areas: a) unplanned and b) planned area (10 times smaller than (a))
UEs' density in all areas of the macrocell is random	UE's density in planned area is 5 times higher than in unplanned area
Only 30% of HBSs employ open-access control policy (residuals are hybrid and closed-access HBSs)	All HBSs employ open-access control policy in the planned area while in the unplanned area the percentage is 30%
HBSs are randomly distributed in the macrocell area	More and sparser HBSs operate in the unplanned area while less and denser HBSs in planned area

In section 3.2.1, various alternatives for the use of the two-fold backhaul were discussed e.g. resource allocation based on the type of service of each UE. In the following scenarios, we consider a static resource allocation where the fUEs are served through the wired backhaul, while the mUEs data is routed through the wireless backhaul. The reference case that is used for comparison is the standard femtocell deployment where only the wired backhaul alternative exists. For the pathloss models, the ones suggested for the urban deployment by the small cell forum specifications in [169] were taken into consideration.

3.5.1 Unplanned femtocell deployment case simulation results

A classic two-tier network, where the macro BS is located at the macrocell center, while HBSs are randomly located, is considered in this scenario. Moreover, mUEs are positioned in random areas of the macrocell following a uniform distribution. Due to the assumed unplanned and users'-initiated manner of HBSs' deployment, a low percentage of open-access femtocells (i.e. 30%) is assumed.

Figure 3.17 presents the results for various sets of UEs/MTs. We denote as cell-edge mUEs those that achieve a spectral efficiency equal or below to 10% of the value of the average mUE in the macrocell. Furthermore, constrained fUEs are those belonging to femtocells where the wired backhaul link is fully/highly-utilized. Hence, when these femtocells are in open-access functionality and associated with a mUE, the wired backhaul is shared among

fUEs and the mUE and thus fUEs experience capacity limitations.

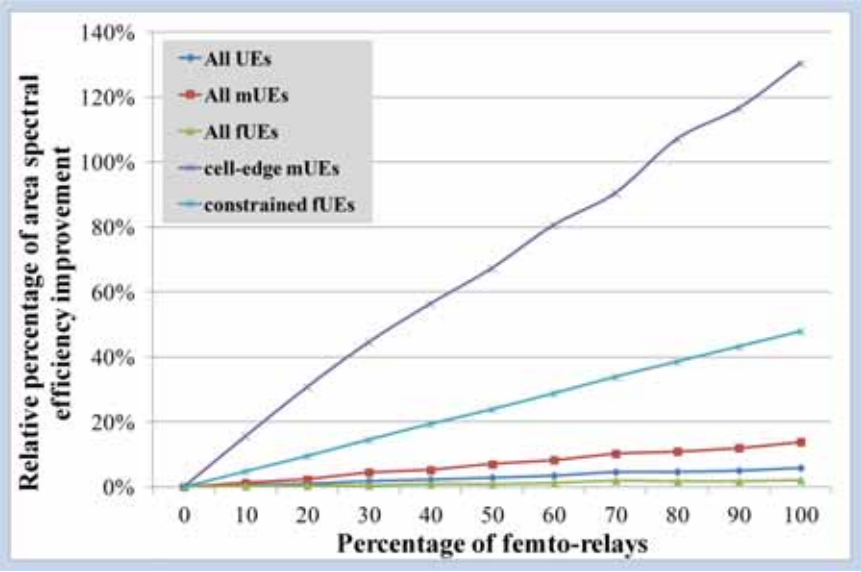

Figure 3.17: Relative improvement of area spectral efficiency for various UE categories for increasing percentage of femto-relays in an unplanned two-tier network

According to simulation results shown in figure 3.17, cell-edge mUEs experience the greatest improvement as femtocells are replaced by femto-relays. This can be easily explained by the fact that cell-edge mUEs have low spectral efficiency values and thus the provisioning of an additional wireless backhaul offers them better throughput levels than these in the standalone shared wired backhaul case. On the other hand, the whole set of fUEs does not experience significant gains from femto-relaying as the backhaul selection module (BSM) presented in section 3.2 selects only the mUEs to be served through the wireless backhaul which provides additional and increased capacity in comparison with the wired backhaul. Thus, we observe that mUEs attain significant greater improvements from femto-relaying. On the contrary, fUEs' gain comes from the wired backhaul offloading by serving the attached mUE(s) wirelessly. In general, fUEs' insignificant improvement is due to the following facts: a) open-access femtocells are the minority of femtocells (30% of the total number of femtocells), b) fully/highly-utilized femtocells are not the majority of the femtocells and c) wired backhaul capacity is lower than the access link capacity which is based on an LTE interface, thus creating an end-to-end bottleneck for the fUEs which do not use the wireless backhaul in the considered scenarios.

Another noteworthy observation is that the average UE's relative improvement curve is closer to the fUEs' one. This is normal as fUEs are assumed to be the majority of all UEs in the network. Another observation one can make is that the average mUE (depicted by the "all mUEs" curve) achieves a significant gain of up to 14% for full femto-relay penetration although quite lower than that of cell-edge mUEs. This gap exists because the average mUE

doesn't suffer from outages and low throughput like cell-edge ones. Moreover, cell-edge mUEs are a small portion of total mUEs (i.e. worst-case mUEs) and thus their gain does not significantly alter the average gain of the whole mUEs set. From the aforementioned, we conclude that femto-relaying is well-suited for fairness improvement as cell-edge mUEs can experience a better QoS. A similar case exists among fUEs and their subset of constrained fUEs. Based on earlier discussions, we conclude that femto-relay concept's contribution mainly targets fUEs associated with fully/highly-utilized femtocells.

3.5.2 Semi-planned femtocell deployment case simulation results

Even though the femto-relay concept can provide promising gains in terms of (area) spectral efficiency for all UE categories, as shown in figure 3.17, its main drawback is that HBSs' owners don't have a clear motivation in adopting femto-relay architecture, which adds a layer of complexity to the plug-and-play nature of currently deployed femtocells. Hence, the femto-relaying success highly depends on its real market penetration rate. Therefore, a semi-planned deployment case is considered (cf. table 3.2), where the femto-relay concept can be widely and more easily adopted in a planned area and where cellular operators' IT professionals will be responsible for the careful femtocell installation and general network planning procedures.

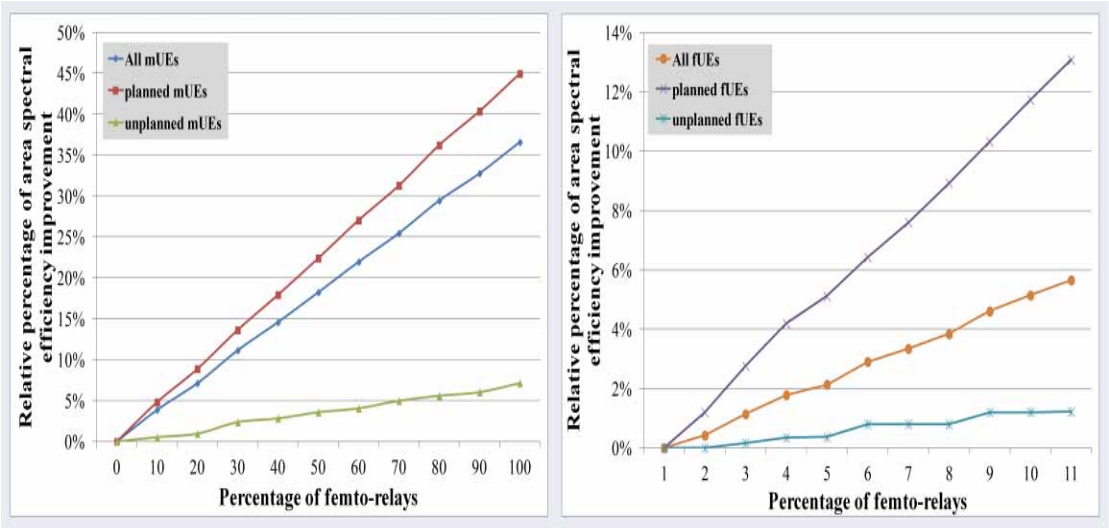

Figure 3.18: Relative improvement of area spectral efficiency for mUEs and fUEs for increasing percentage of femto-relays in a semi-planned 2-tier network

In figure 3.18, depicted planned UEs are the ones served within the planned region, whereas, unplanned UEs are those located in a cell region that has the same parameters as the ones that were applied in the unplanned femtocell deployment case. Planned mUEs increase their area spectral efficiency up to 45% thanks to the regular use of wireless backhaul alternative offered by femto-relays. On the other hand, unplanned mUEs are located in the cell region, which is dominated by closed-access femtocells, achieving an improvement of up to 8% due

to the low level of femtocell association and the smaller density of mUEs. Furthermore, we observe that the average mUE curve is closer to the planned mUEs' one. This is accredited to the larger number of planned mUEs compared to unplanned ones. In figure 3.18(b), one may see that the planned fUEs achieve significant improvements as femto-relay penetration increases. As all the femtocells adopt an open-access policy in the planned area, cooperation with the mUEs would degrade the fUEs communication as the mUEs consume a part of their wired capacity. In this case, the addition of the wireless backhaul reserves the wired backhaul capacity only for the fUEs. On the contrary, unplanned fUEs are mostly associated with closed-access femtocells and they rarely share their backhaul with low-density mUEs.

A final observation concerns the comparison of femto-relay area spectral efficiency gains for the average mUE and fUE between the two network deployment cases. It is obvious that in the semi-planned case, where the femto-relay real market penetration rate can be considerably greater than in the classic unplanned case, the proposed concept's contributions can be even more promising, when seen from an overall 4G HetNet's system perspective. Conclusively, this finding is completely aligned with the cellular operators' strategic objective to operate femtocells in open-access mode under all cellular user density circumstances, in order to be able to both maximize their revenues and provide optimal QoS to their customers.

3.6 Summary

In this chapter, the femto-relay concept was introduced and studied, which enhances the currently deployed 4G HetNet environment in various ways. An overview of the main challenges that degrade the performance of two-tier networks has been provided, that is to say interference management and limited backhaul issues. In continuity, four main functionalities of the proposed femto-relay concept were presented. Furthermore, two context aware frameworks for the femtocells' efficient integration in existing IP and cellular infrastructures were proposed and evaluated covering both uplink and downlink cases. Two realistic network deployment cases denoted as unplanned and semi-planned were considered and for these settings, the performance of femto-relaying was evaluated in terms of area spectral efficiency enhancements. Simulation results have also shown that the overall performance of a HetNet environment can be leveraged in terms of QoS requirements for various MT/UE groups, energy saving and data rate enhancement.

CHAPTER 4

CONTEXT AWARE RESOURCE MANAGEMENT FOR MOBILE AND FIXED NETWORKING SYSTEMS' CONVERGENCE

In this chapter, context aware resource management frameworks being applicable for mobile and fixed networking systems' convergence are studied. According to figure 1.1 and regarding the three main architectural pillars of the current thesis, the scope of the research being undertaken in this chapter refers to the respective outlined area in the middle of the figure (i.e. mobile and fixed networking systems' convergence). More specifically, this term's main concepts are introduced accompanied by state-of-the-art works found in the international literature. In section 4.2, the novelty features of the proposed integrated services router (ISR) concept are introduced, while some other related real-market products are described such as small cell gateway, machine type communication (MTC) gateway, etc. In sections 4.3-4.5, three decision making schemes (i.e. DSAC, ISAC and CABM) are proposed, which can reside at ISRs providing functionalities such as integrated QoS provisioning, admission control and backhaul management. For all proposed schemes, problem formulation, algorithm description and performance evaluation results are provided. The final conclusions are summed up in section 4.6.

4.1 Introduction

In the Future Internet (FI) environment, it is expected that a variety of wireless networks will coexist and collaborate transparently. Furthermore, as the services provided by the various networks are increasingly converging, it will be possible for the users to access the same service with the same QoS through different networks [49] [186] [187]. For example, despite the fact that cellular networks were primarily designed to cater for voice services with mobility, nowadays all types of services can be supported by 3G and beyond cellular systems. Similarly, Wireless Local Area Networks (WLANs) were mainly developed as an extension of terrestrial LANs in order to serve burst-traffic data services and nowadays VoIP services served by WiFi hotspots have also become very popular. This situation is beneficial for both the users and the service providers as it serves the "Always Best Connected" (ABC) concept [12], while it can also be utilized by load balancing mechanisms in order to transfer traffic load between the interworking networks [188]. The FI vision is also expected to extend the ABC notion and include use cases according to which the FI end user will attain any service on a single intelligent device (e.g. smartphone, tablet, etc) using any available network within a heterogeneous network environment consisting of open, cognitive and collaborative

wireless and wireline networks. Towards this direction, Heterogeneous Network (HetNet) deployments, proposed by 3GPP LTE-A standardization group, are expected to become a reality in the next few years providing thus efficient ways to deal with continuously growing traffic demand [30]. Furthermore, FI is expected to sustain a huge number of devices, many orders of magnitude higher than state-of-the-art network architectures and handle larger and unconventional information flows. Machine Type communications (MTC) will play a critical role in this so called Internet of Things (IoT) evolution and thus both architectural and algorithmic enhancements to existing converged mobile and fixed network infrastructures are needed [189] [190].

So far, existing resource management solutions consider that the main capacity bottleneck of cellular networks is mainly on the air/radio interface [191]. However, cellular network capacity limitations due to backhaul capacity shortage should not be underestimated in some small cell and 4G HetNet network deployments such as the one considered in the current thesis' context (cf. figure 1.1). Indeed, it is expected that backhaul connections to small/femto cells will be forced to carry additional traffic and thus Internet Service Providers (ISPs) will end up responsible for a large portion of the total mobile data traffic incurring many doubts about QoS/QoE provisioning issues of the delivered services [168] [192]. For example, the performance of a Femtocell Access Point (FAP) depends on the quality of the underlying backhaul, as it is likely that the available bandwidth link is below the FAP's peak data rate and hence the bottleneck can be occurred at the xDSL router (i.e. see proposed ISR concept in section 4.2). Conclusively, one may assess that the quality of mobile services provided by ISRs depends not only on the radio link quality from the MTs/UEs to their associated access point interfaces, but also on the level of congestion at the corresponding backhaul link [193]. In this chapter, resource management schemes are proposed, which are aware of the 4G HetNet environment conditions and the available backhaul resources and thus are considered as "context-aware" in the current thesis framework.

4.1.1 Data traffic aggregation points

In contrast with the widely known cellular networks, which consist of a dedicated terrestrial backbone, femtocell is an emerging technology which uses the IP backhaul network along with small-size base stations located indoors [31]. Hence, the advent of femtocells added a new perspective on the ABC concept as a femtocell may operate together with other wireless or wired networks, which however often share the same backhaul capacity [194]. Therefore, the increase of the traffic load in one network reduces the fraction of the capacity that can be utilized by the other networks. Consequently, while a user is practically able to initiate the same service through different interfaces, he is allocated capacity from the same capacity pool. Taking into consideration the fact that xDSL backhaul capacity is physically limited, the

QoS being offered can be diminished in terms of user's satisfaction level (call blocking probabilities, throughput, etc.) [195]. Moreover, despite the convergence of services, each network has a different origin and appeals better to a different kind of users while is able to maintain different QoS levels. While that justifies the presence of various networks covering the same geographical area at the same time, it may also cause problems when these networks have to interwork and share the same backhaul capacity. Specifically, while the femtocell inherits the QoS mechanisms of cellular networks and is able to provide a reliable Call Admission Control (CAC), this does not apply to the IP-based networks and this fact may drastically affect the performance of the femtocell.

Nowadays, there is a emerging trend in having various types of ISRs (e.g. small cell gateways, MTC gateways, etc), which consist an aggregation point for mobile data traffic. As mobile data traffic aggregation points, one may consider 4G HetNet entities, which accumulate mobile data traffic and route it through the Internet to residual cellular and IP fixed infrastructures. Due to the fact that 4G HetNet/small cell/MTC traffic is expected to experience a continuous increase during the upcoming years, backhaul-aware resource management challenges should be timely addressed. In this chapter, all proposed context aware resource management schemes are residing at mobile data traffic aggregation network entities (i.e. ISRs), whose names differ according to the problem formulation being assumed (i.e. integrated xDSL router, small cell gateway and MTC gateway for sections 4.3, 4.4 and 4.5 respectively).

4.1.2 Integrated services QoS provisioning

Based on the above descriptions, there are mobile data traffic aggregation points, in which efficient network resources management procedures should take place. Several QoS considerations related to the femtocell network architecture have arisen and deal with the question how can backhaul network provide acceptable QoS [194]. The existence of net heterogeneity poses a serious concern, especially in cases, where the wireline backhaul provider is not in tight strategic relationship with the cellular operator. Moreover, as FAPs are part of a continuously changing 4G HetNet environment, femtocells may experience difficulty transferring even low bandwidth services due to wired backhaul limitations. The limitations of xDSL backhaul capacity is a vital issue that needs to be addressed. Recent 3GPP standardization activities assert that it shall be possible for the network to set different criteria for access control in a hybrid cell for closed subscriber group (CSG) and non-CSG members [196]. These different criteria incur conflicting results and this problem can be tackled by specific admission control policies as addressed by the three schemes proposed in this chapter.

4.1.3 Related work overview

During the last years, a noticeable increment in the demand of indoor voice and data traffic had been observed. Hence, how to provide good indoor coverage, in particular, for resource consuming data services, has become a major challenge for operators [31]. This was the time when femtocell concept came in the market and research foreground in order to deliver high quality of service (QoS) to users at home, at work and even at public places of special interest. This high QoS includes reliable coverage and high data rates while the cost per bit for delivery is reduced [197]. Thus, it is well recognized that femtocells can bring a lot of advantages for both operators and subscribers providing a win-win market situation. Collaboration aspects with already existing WLANs are being investigated and it is shown that IEEE 802.11 and femtocell technologies have many complementary features. For instance, cellular-WiFi convergence concept refers to simultaneous provisioning of both cellular and WiFi services via a single integrated network. According to a related recently published white paper by Small Cell Forum [198], some benefits of Integrated Femto and WiFi (IFW) networks in various 4G HetNet and small cell deployment scenarios are: a) compared to WiFi only scenario, IFW network brings all mobile network operator's (MNO) services and inexpensive plain internet access to all 3G & WiFi devices, b) simplify LIPA and SIPTO operations for WiFi+3G devices, (e.g. using WiFi for LIPA and 3G for MNO services) [199], c) reduce interference between femto and macro-cellular networks by offloading to unlicensed WiFi radio access, whenever possible, d) enable integrated QoS provisioning in the bandwidth management processes (e.g. aggregation of IP flows) across the IFW network radio links, e) enable integrated services across the 3G+WiFi devices. As the mainstream real-market trend tends to have both femto and WiFi technologies residing in the same IFW product (e.g. some initial product releases are already available by Cisco, Alcatel-Lucent, Netgear, NEC, etc), an integrated QoS provisioning approach has to be adopted for these, so called, small cell gateways. These small cell gateways may be applicable for diversified geographical areas, in which IFW networks are deployed. In [32], an extensive study has been made providing explicit differences and similarities between residential, enterprise and public real-market network deployment scenarios, while it is stated that proposed resource management and QoS provisioning modules have to be flexible enough to cope with various small cell deployments heterogeneities (see more in 4.2.1). The same rationale is proposed in [190] for machine-to-machine (M2M) communications outlining the need for efficient wired backhaul traffic management by a QoS provisioning module residing at a MTC gateway (MTCG). Moreover, in works such as [200], it is stated that femtocells provide an ideal solution for Fixed Mobile Convergence (FMC). In this case, operators can bring all potential wireless and wired users under their realm over such an integrated environment, providing thus users with the benefits of ubiquitous connectivity.

Referring to recent research issues related to the emerging femtocell concept, several papers have been published dealing with interference management in two-tier networks and how this problem can be tackled along with the selection of the appropriate access control mechanism [30] [170] [171] [173] [201]. There are three access control mechanisms referred in the literature: a) closed access, where only a predefined subset of users (Closed Subscriber Group - CSG) can connect to the FAP, b) open access, where all subscribers of the operator have the right to make use of any FAP and c) hybrid access, where part of the FAP resources is operated in open access mode while the remaining follow a CSG approach. In general, the adoption of a closed access mechanism leads to higher levels of femtocell subscribers' satisfaction but creates high levels of overall system interference in the proximities of the FAP. On the other hand, open access provides a better overall network performance in terms of throughput and utilization but the signaling overhead due to often handover procedures is increased and the CSG's satisfaction level decreases abruptly. Consequently, it is accepted that hybrid access mechanisms seem to be the best solution, as they can be seen as a trade-off between the first two approaches. However, much discussion can be made about the amount of shared resources, which have to be carefully tuned according to the CSG and the non-CSG members' profile and the environment in which the whole system architecture is deployed.

Several QoS considerations related to the femtocell network architecture arise and have to deal with the question how can backhaul network provide acceptable QoS [194]. Towards finding ways to tackle this problem and taking into account that FAPs use the incorporated HSPA scheduler, the FAP's QoS module has to somehow cooperate with the corresponding QoS module of the main router, which interconnects the various network interfaces with the IP backbone. Another factor, which can diminish the users' satisfaction level, is the limitations of the xDSL backhaul capacity. In [195], this problem is tackled by proposing SLA negotiations using bandwidth broker in order to reserve sufficient bandwidth for femtocell users. The need for existence of a QoS scheme, which has to be carefully designed in order to preserve the desirable QoS of mobile traffic as well as that of fixed traffic, is also outlined in [202].

Recent 3GPP standardization activities assert that efficient admission control policies should be developed for integrated services QoS provisioning [181] [196]. Generally, there are many challenges in call admission control (CAC) design for heterogeneous wireless networks [121]. More specifically, CAC schemes should be designed to support different types of user groups and services with corresponding QoS requirements. Interoperability with DiffServ-based IP networks should also be supported. Moreover, adaptive bandwidth allocation is also necessary in order to improve utilization of wireless network resources. Proposed schemes in sections 4.3-4.5 can be considered as complementary to [195], which presents a scheme for dynamic SLA negotiation between the ISP provider and the femtocell operator, aiming to

tackle the problem of xDSL's varying capacity. In general, this chapter's proposed schemes belong to the general family of hybrid partitioning policies [153] [203] and can be classified within the family of admission control schemes designed for the integration of heterogeneous networks such as [204], [205] and [206]. However, such schemes differ with the proposed ones in this thesis as: a) they consider significantly different system models, which in most cases include networks that have their own separate links to the core network or b) they support basic QoS features that are not compatible with the strict QoS requirements of a femtocell.

4.2 The Integrated Services Router (ISR) Concept

As already mentioned, there is a market trend in having integrated service routers (e.g. integrated WiFi/FAP routers, small cell gateways, machine-to-machine gateways, etc), which consist an aggregation point for mobile data traffic. As a result, integrated QoS provisioning and efficient resource management solutions are needed in order to achieve an efficient integration of small cells in existing IP and cellular infrastructures. The idea of combining WiFi, femtocell, router and DSL modem in a single box was first introduced by NETGEAR in 2008 and since then many vendors and operators have encouraged their fixed line and mobile teams to combine their strategies, breaking down the silos between them.

In this thesis and according to figures 1.1 and 3.6, the ISR entity can be considered as an ideal network entity that can take into consideration both radio and backhaul resource requirements and possible bottlenecks and thus appropriately coordinate the overall resource management procedure for converged mobile and fixed networking systems. Conclusively, the proposed context aware resource management schemes presented in sections 4.3-4.5 are assumed to be residing in ISR-like entities.

4.2.1 Real market network deployment scenarios

In this section, the different use cases described in [196] are combined and classified into three main real market network deployment scenarios. The conflicting issues described in the related work overview section are also mapped appropriately into each scenario being assumed. The main key performance indicators that diversify the following scenarios (see also table 4.1) are: a) the type of FAPs being deployed, b) the type of services being prioritized, c) the type of the environment context being assumed, d) the definition of CSG and non-CSG sets of users for both the FAPs and the WiFi APs, e) the definition of priority groups within the CSG list, and f) the extent of wired network usage.

Table 4.1: Key performance indicators for the 3 real market network deployment scenarios

	Home User Scenario	**Enterprise Scenario**	**Public Access Scenario**
Type of FAPs	Class 1	Class 2	Class 3
Services being prioritized	Live/streaming	Interactive/live	Streaming/interactive/live
Environment context	urban vs. rural (sub-urban) use case	building vs. campus use case	Metro/railway stations, airports, stadiums, shopping malls
FAP's/WiFi AP's access control mechanisms	hybrid (large CSG)/closed	hybrid/hybrid	hybrid (small CSG)/open
CSG list priority groups	N/A	Special guests & stuff with higher/lower priority	N/A
Wired network usage	medium	relatively high	low

Regarding (a), there are three main classes of FAPs [197]. Class 1 is the currently best known model, which can support four simultaneous voice channels plus data services and is installed by the end user at his/her home. Class 2 FAPs transmit in higher power, supporting longer range and more concurrent connections. Finally, class 3 FAPs can cover even longer range, support up to 32 concurrent connections and are often deployed in built-up areas and crowded public places. The services being offered are the common voice/video calls, streaming/video on demand services, web browsing, file downloading, e-mail etc. and some emerging ones such as mobile TV, IPTV, internet gaming, online social networking applications etc. In table 4.1, the types of services, which have to be prioritized for each assuming scenario, are given in a sententious manner. For example, in home user scenario, live and streaming calls have the highest priority while in enterprise scenario our proposed scheme has to take special care of the increased demand of interactive calls. In any case, our proposed scheme has to take into consideration the effect of traffic dynamics on the resource management policy, regarding the fact that depending on each service requirements, different amounts of resources may be required on different timescales [207]. The area in which the overall network architecture is deployed, can be sparse or dense populated, affecting hence directly the system performance. For example, urban use case (home user scenario) and building use case (enterprise scenario)

refer to dense populated areas, while rural and campus use cases refer to sparse populated ones. The definition of the user priority groups in each scenario is also very important and is related with the access control mechanism being adopted [173]. In table 4.1, it is shown that, generally, hybrid access control mechanism is adopted for FAPs allowing though different amount of resources to follow a CSG approach for each assuming scenario. Regarding WiFi APs, they can be closed, open (public hotspots) or operate in hybrid mode. Finally, the increase of wired network usage can severely reduce the fraction of capacity that can be utilized by the other wireless networks and can thus affect the overall system performance, too.

All the features referred above related to key performance indicators for the three different real market network deployment scenarios are summarized in table 4.1. In the rest of this section, the scenarios are described in a more detailed fashion. Thus, this section provides important feedback to the design and the implementation rationale points of the proposed DSAC, ISAC and CABM schemes presented in sections 4.3-4.5.

4.2.1.1 Home user scenario

The first scenario describes the situation where one random user sets up his/her own network at his/her own residence. In this scenario, a small number of users (class 1 FAPs), large bandwidth and time consuming services are assumed. The paying home user wants to enjoy a large variety of services perceiving high QoS at the minimum cost. Therefore, he/she wants to enjoy high data rate delay tolerant services (file downloading, streaming videos, internet games etc) from his Wi-Fi AP and at the same time to be able to accomplish delay sensitive services (voice calls, live video calls, mobile TV, etc.) from the existing residential FAP in the most effective way.

From the single paying home user's perspective, all the available backhaul network resources should be kept for the group of users (CSG group) he has manually determined during the initial FAP setup. On the other hand, the network operators would prefer an open access deployment since this provides an inexpensive way to expand their network coverage and capacity capabilities. The truth about the feasible real market implementations is somewhere between the two pre-referred perspectives. This means that the home user can substantially benefit from assigning bandwidth (from his/her own FAP) to passing by macrocell users and his guests remaining at his house for a couple of hours. These users belong to non-CSG group and in case they are connected to the overlaying macrocell, they can cause severe QoS degradation problems to the CSG users' perceived QoS due to the uplink interference they incur (read more in chapter 3). Consequently, we assume a hybrid access mechanism assigning high priority to CSG users of femtocell. The grade of priority can be adjusted according to the use case being assumed. More specifically, we assume 2 different use cases

in this scenario. The point that diversifies them is the population density. Rural and sub-urban areas, which are sparse populated, comprise the first use case while the second one refers to dense populated urban areas. In the former use case, there are only a few non-CSG users and thus a small portion of the available bandwidth should be assigned to them. On the contrary, the proposed context aware resource management schemes have to be flexible enough to cope with the increased bandwidth requirements of non-CSG users in the case of dense-populated areas without degrading the desirable perceived QoS of CSG users services. Wi-Fi is being assumed in closed access by all means in this scenario.

For both use cases, the proposed schemes have to assign the highest priority to delay sensitive real-time services requested by CSG users. That is, the home user's CSG list should (almost) never come to an "overload state". The second main goal is to achieve an optimal system utilization rate, which will allow a maximum number of non-CSG users to be served by the system.

4.2.1.2 Enterprise scenario

The places where the enterprise scenario can be applied can be enterprises, organizations, ministries, conference places, research centers etc being accommodated in entire buildings (use case A) or in organized campuses (use case B). A larger number of users (e.g. 8-16 concurrent connections) can be served. The whole network architecture installation study can be made either by the each time organization itself or by its ICT subcontractors in order an optimal network coverage planning to be made. In this type of environments, apart from the pre-existing WiFi APs, there are wireline infrastructures, too. For example, employees who work permanently in the building's or campus' premises may use the reliable wired line as they are spending most of their time in front of their desktop PCs. On the other hand, they also often have to move to other offices along with their mobile devices in order to cooperate with their internal colleagues or to attend in meetings with their guest colleagues.

From the above, it can be inferred that the main particularity of this scenario lies in the fact that the borders between the CSG and the non-CSG lists are not clearly specified. For example, a guest entering the building or campus can be a member of both lists and this depends on how often he/she visits the specific place. If he/she attends a conference or a meeting for a couple of days and then leaves, then this user should be a member of the non-CSG list. On the other hand, if he/she is an often guest colleague, then this user can be added in the CSG list by the system administrator. In the latter case, more than one subgroup can be defined in the CSG list, each one of them having different priorities. WiFi APs are assumed to be open in places, where there are a lot of non-CSG users (reception or meeting rooms) and closed (only for stuff needs) in any other case.

We assume two different use cases: A) the enterprise-building and B) the enterprise-campus.

The main points that diversify these use cases are: a) the population density of the area where the assumed network architecture is deployed, b) the synthesis of CSG and non-CSG lists, and c) the services being prioritized. Regarding the enterprise-building use case, CSG users are the employees and any other regular guests who are registered in the CSG list. The non-CSG list includes the remainder guests and the citizens who remain in the assumed area for a couple of hours and then leave. These users want to experience adequately high QoS during their stay in the building without influencing the perceived QoS of the CSG users. Therefore, the bandwidth reserved for non-CSG users is sensibly larger than the corresponding one in home user scenario. In enterprise-campus use case, non-CSG users request for even higher bandwidth. The type of services, whose request increases abruptly in both use cases, are the so called data calls such as web browsing, e-mail, file downloading, instant messaging etc. However, conventional voice calls' QoS has to be guaranteed by all means, even during working hours when the system utilization is high. Consequently, the proposed schemes must be flexible enough to cope with a large variety of different situations, which are continuously changing.

4.2.1.3 Public access scenario

Public access scenario can be applied in places such as metro stations, railway stations, airports, shopping malls, stadiums etc. That is, crowded public hotspot areas where the use of FAPs can alleviate the heavy macrocell traffic in the most remarkable way. Another main advantage in such kind of network deployments is that MTs will have the chance to connect to closer NodeBs (i.e. FAPs) optimizing the system's coverage. Moreover, MTs are consuming much less power. This is very important in the places being assumed in this scenario, where available power supply cannot be easily found.

Class 3 FAPs are deployed and thus at least 16 concurrent connections can be supported by each one of them. The vast majority of users belong in the non-CSG list. That is, we substantially consider an open access control mechanism with some exceptions. These are the employees who are working in the assuming coverage area. The largest priority should be assigned to these CSG members. The calls of these users are strictly specified by the rules defined by their superiors. WiFi APs are supposed to be open (public hotspots). Interference problems are assumed to be manageable in this kind of environments as the operator's professional staff has the entire responsibility of the network planning optimization.

In this scenario, an abrupt increase in video streaming services is observed. For instance, travellers waiting for their flights at the airport, individuals waiting for their wives/husbands and friends at a shopping mall, fans willing to see an instant replay of the game they are watching at a stadium etc., request for video streaming services at a very high percentage. Moreover, in such places a new trend aiming to offer personalized services via video

streaming guides and advertisements to users is emerging. Therefore, the particularity of this scenario is that high resource consuming services are assumed and therefore these have to be treated with special care by the proposed schemes, without violating the basic principles described earlier in this section.

4.2.2 Small cell gateways

Mobile subscribers want access to the network at home, work, hotspots, and everywhere in between. This requires mobile operators to expand their service offerings over multiple new access networks, such as broadband DSL, fiber to the home, or cable broadband networks, using small cell technologies such as femtocells, Wi-Fi and other WPAN technologies. Hence, operators can expand high quality cellular coverage beyond the reach of their existing footprint to fixed line customers at home, assure QoS to mobile devices, and seamlessly move traffic from macrocells to small cells, increasing spectrum and capacity utilization efficiency. A small cell gateway is a single network entity integrating FAP, Wi-Fi AP and other wireless access technologies (e.g. WPANs). This solution is introduced to aggregate a large number of LTE small cell traffic backhauled over fixed line broadband links such as cable modem, DSL or fibre, thus enabling efficient connection to the operator's core network. Some main small cell gateway's advantages are the following: a) provides an exceptional customer experience, which helps to increase revenue and foster customer loyalty, b) reduces the number of components in the network, decreasing the number of potential points of failure and promoting a lower capital and operating expense model, c) gives operators the security and freedom to choose the right access application for their specific needs, without the requirement to design access strategy around the limitations of multiple diverse products, d) provides the operator with the same transparent service experience required to deliver excellent user experiences and delivers the intelligence necessary to promote new sources of revenue, e) provides end users with a seamless experience between macro and small cell networks (including WiFi APs, WPAN technologies, etc).

From the above-mentioned advantages of small cell gateways, it is expected that this type of devices will prevail in the telecommunications' market in the next decade. As a matter of fact, many real market products have come in the foreground during the last couple of years. Big industrial players such as Cisco [208], Netgear [209], Alcatel-Lucent [210], NEC [211], etc are focusing on mobile and fixed networking systems' convergence and consider small cell gateway as key network entity towards the realization of this vision, while other smaller players like Astri [212], BTI Wireless [213], Lever technology group [214] and other emerging SMEs and spin-off/out companies are trying to get their own market share.

4.2.3 Machine type communication gateways

As Future Internet (FI) is expected to sustain a huge number of devices realizing the Internet of Things vision, machine type communications (MTC) will play a critical role in this evolution. The mobile data traffic that is generated by MTC devices (MTCDs) is also expected to experience an exponential increase in the next decade and thus network aggregation points called MTC gateways will come in the telecommunications' market foreground [28].

There are several machine-to-machine (M2M) access network architecture variants ranging from: a) wireless cellular solutions, which rely on wide coverage (i.e. GSM, UMTS, LTE, etc) to b) purely embedded short-range solutions relying on cheap and energy-efficient deployments (i.e. low power WiFi, Bluetooth, Zigbee, etc). Recently, hybrid architectural solutions providing good trade-offs among price, range, rate, scale and power were introduced [215]. MTCG is a critical architectural element for such kind of hybrid solutions as there is a need for mobile data aggregation at some point, preferably as close to MTCDs as possible. The MTCG facilitates communications among a great number of MTCDs and provides a connection to a backhaul that reaches the Internet, while it appears to the cellular network as a client that competes with cellular users but also with other MTCGs [216].

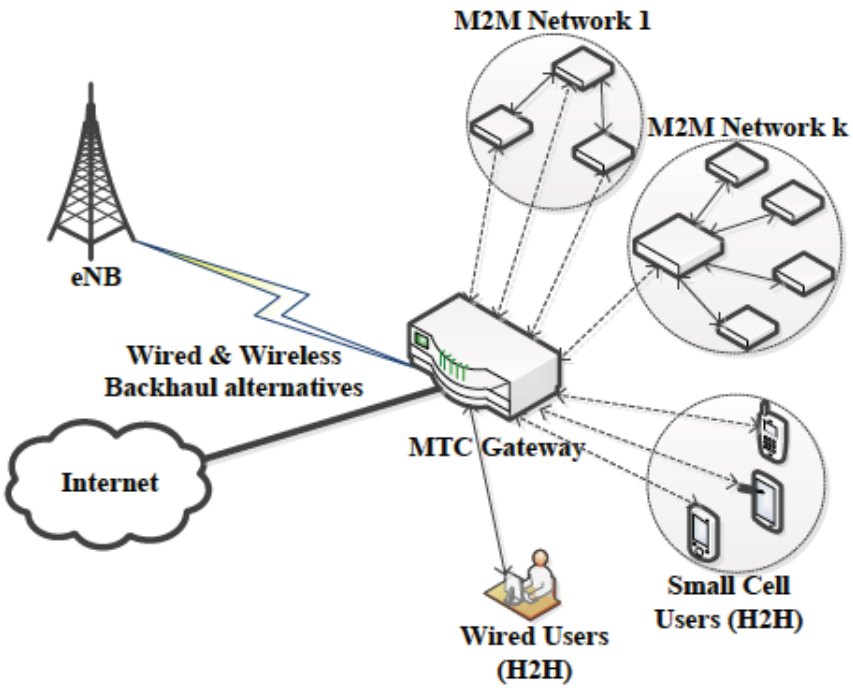

Figure 4.1: The MTC Gateway

In figure 4.1, the assumed hybrid M2M access network architecture is depicted. The cornerstone of the assumed system model is the MTCG, which plays a key role as an aggregation point in bringing the installed short-range sensors/MTCDs online and providing interworking with different wireless access technologies [217]. MTCDs can directly connect

to MTCG via wired/wireless interfaces (cf. M2M network 1) or get indirectly connected through a master MTCD/cluster head (cf. M2M network k). At the same time, small cell/desktop users can make use of the MTCG's cellular/wired interface and thus generate the so called human-to-human (H2H) traffic. This deployment setting has many similarities with the one, which includes a small cell gateway or an integrated femtocell/WiFi access point instead of a MTCG (section 4.2.2). The main difference is a MTCG can further improve the network architecture to accommodate novel MTC service requirements without sacrificing the current H2H services' QoS. Moreover, both wired and wireless backhaul alternatives are provided to MTCG. Wired backhaul may be preferred in residential or even enterprise scenarios [218], while wireless backhaul is surely preferable in public access scenarios (see section 4.2.1.3).

4.3 Proposed Dynamic Service Admission Control (DSAC) Scheme

The system model depicted in figure 3.6 is assumed for this section. The analysis of some distinctive network deployment scenarios in the previous section renders obvious that almost in all cases a femtocell has to operate in hybrid access mode, assigning thus a part of its capacity to CSG users and the residual part to non-CSG users. Furthermore, a coexisting WiFi AP could also serve a closed group of users or may serve any user that comes into its range. Finally, wireline network users may also claim their share of the same backhaul capacity. Therefore, the employed QoS policy should be user-centered and flexible enough so as to be easily adapted to various network deployment scenarios.

Consequently, the objective is to propose a framework in which a unified service admission control policy can be applied. In this framework, the users are not classified according to their network interface but instead according to the user group that they belong and the type of service they request. Each user group may be treated differently by the admission control process allowing thus the network administrator to apply a QoS policy tailored to a specific traffic load composition. While proposed DSAC scheme uses capacity partitioning in order to provide differentiated QoS, however it provides better performance compared to classic partitioning schemes as it allows the dynamic utilization of the partitioned capacity.

4.3.1 Service requests' and user groups' classification

4.3.1.1 Service class integration

The first step in providing a unified service admission control policy is to group the various services provided by the three network interfaces (i.e. FAP, WiFi AP and wireline network) into six common Service Classes (SCs). Thus, the admission of the service calls will be based on the service class they belong and not on the network interface that the user employs. The basic criterion in order to classify a service to a specific service class is its QoS and rate

requirements. Thus, the members of each service class have similar characteristics and consequently it is easier to be treated equally by the admission control process. Therefore, we adopt the following mapping of the services [219] as it is shown at table 4.2.

Table 4.2: Service classes of the integrated network

Service Class	Intergrated Network
SC 1	Broadband - conversational
SC 2	Broadband - streaming
SC 3	Narrowband - conversational
SC 4	Narrowband - streaming
SC 5	Interactive
SC 6	Background

The first two SCs mostly include WiFi and wireline services while SCs from 3 to 6 correspond to the four UMTS service classes (conversational, streaming, interactive and background [181]) including however similar services from the other two network interfaces.

4.3.1.2 Defining user groups

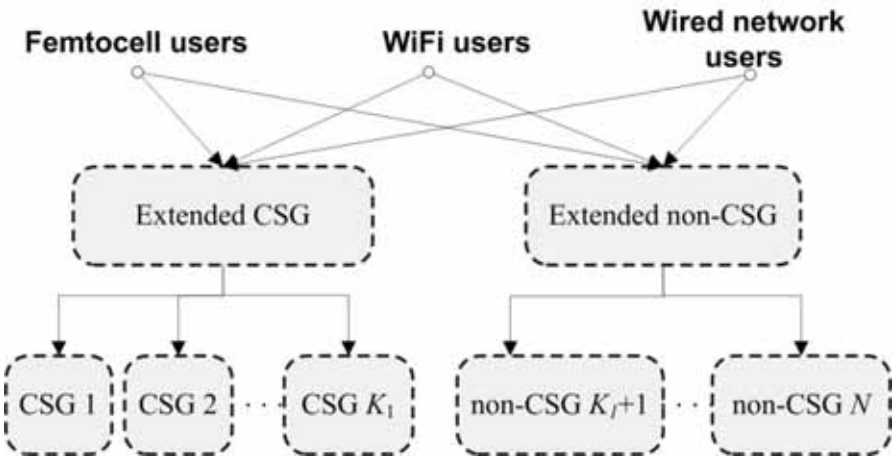

Figure 4.2: Extended CSG and non-CSG user groups

Beside the classification of the service calls into SCs, it is essential to differentiate users into user groups. Thus, we can provide differentiated QoS according to the group that a user belongs. Based on [196], we primarily follow the femtocell approach by defining two main user groups, which are mapped to the CSG and the non-CSG. However, as we refer to an integrated network, we extend these groups so as to include all the network users. Therefore, internal WiFi and wireline users have also to be registered, through their MAC address, to the CSG while all the unregistered users will be treated as members of the non-CSG group.

Moreover, each of the main user groups may be further divided to two or more subgroups if, according to the network deployment settings, there is a requirement to differentiate the QoS and access privileges of the users within the same main group.

4.3.2 Backhaul capacity partitioning

Capacity partitioning is a well-known concept that can be used in order to provide a predetermined QoS, in terms of acceptance rate, to the users. We utilize this concept by dividing the available backhaul capacity into P partitions. The number of the partitions is one more than the number of SCs. The first six partitions reserve capacity for the respective six SCs and their size is defined so as to serve their expected peak traffic load without exceeding a predefined blocking probability. On the other hand, the last partition is used for serving calls without QoS guarantees, as its size is determined by subtracting the summation of the capacities of the first six partitions from the available backhaul capacity.

4.3.2.1 Service calls mapping

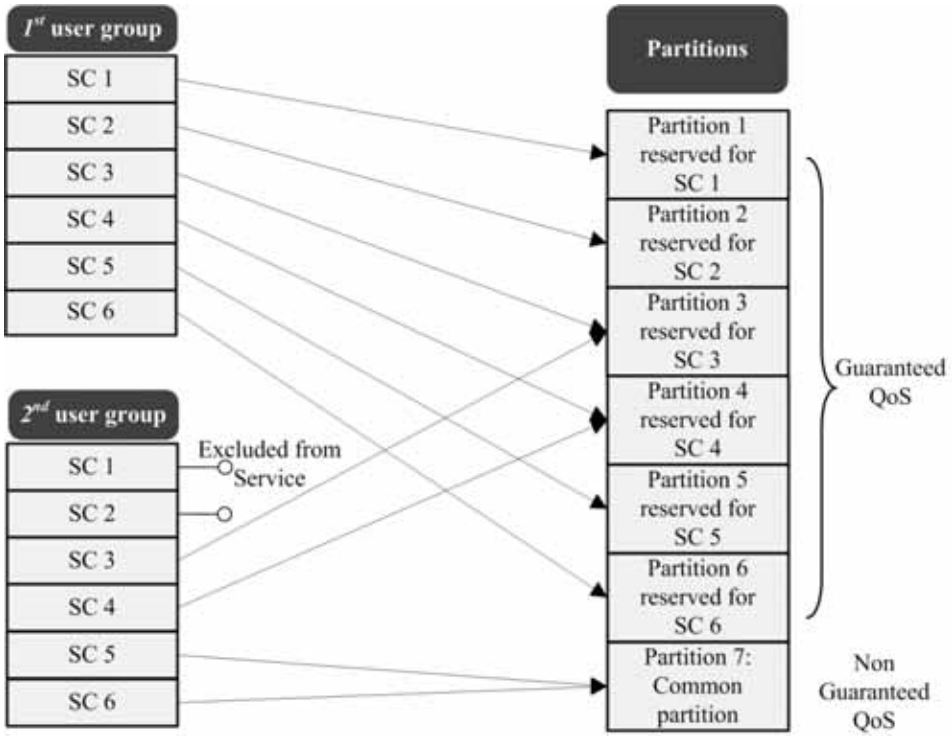

Figure 4.3: Example of mapping the calls of a user group to partitions

As previously discussed, each incoming service call is characterized by the service class SC_i ($1 \leq i \leq 6$) it belongs and the user group j ($1 \leq i \leq N$), which is initiated from. All the service calls with the same characteristics (e.g. SC and user group) can be: a) mapped to be served by the respective partition i under a predefined maximum blocking probability or, b) they can be mapped to be served by the last partition without QoS guarantees or, c) they can be mapped to

zero, which indicates that they will not to be accepted to the system. An example of service call mapping for two user groups is shown in figure 4.3. For the 1st user group, the mapping is straightforward as each SC is mapped to its respective partition. In contrast, for the 2nd user group the broadband SCs (1, 2) are excluded from service and therefore cannot be utilized by the users of the specific group. However, the same users are permitted to access narrowband conversational and streaming services (SCs 3 and 4), which are normally served by the respective partitions. Furthermore, SCs 5 and 6 are mapped to be served at the last partition without QoS guarantees.

4.3.2.2 Structure of a partition

While the partitioning of the backhaul capacity is effective in providing a guaranteed level of QoS under peak traffic, it also causes underutilization when the actual traffic load composition is varying compared to the expected traffic load composition. For example, if the traffic load aimed for a specific partition is temporarily higher than that it can handle then the blocking rate of the respective service class will exceed the required threshold. Furthermore, if the rest of the partitions are low or moderately occupied, then the overall system capacity will be underutilized.

In order to confront the short-term variations of the traffic load composition and increase the utilization of the system, each partition is allowed to accept "external" service calls (i.e. calls which were initially aimed to be served by other partitions). However, this has to be performed in a controlled manner in order to prevent the flooding of the partitions with external calls. Hence, for each partition i we define two areas, the commonly shared area S_i and the reserved area B_i. While an "external" service call can be placed only at the commonly shared area, a "native" service call (i.e. a call which is mapped to be served by the specific partition) can be placed in any of the defined areas. If P_i is the total capacity of partition i, then the reserved area B_i is defined as:

$$B_i = b_i \cdot P_i \qquad (4.1)$$

where $b_i \in [0,1]$ is the respective reservation factor. Consequently, the commonly shared area of the partition can be obtained by subtracting B_i from P_i:

$$S_i = P_i - B_i = (1 - b_i) \cdot P_i \qquad (4.2)$$

As the reservation factor b_i increases, the S_i area decreases and becomes zero when b_i is equal to 1. In this case, the partition does not accept any external service calls. Conversely, when b_i decreases, the S_i area increases and becomes equal to total capacity of the partition when b_i is equal to 0. Then, the partition accepts equally external and native service calls performing thus as a common capacity pool.

4.3.2.3 Capacity allocation process

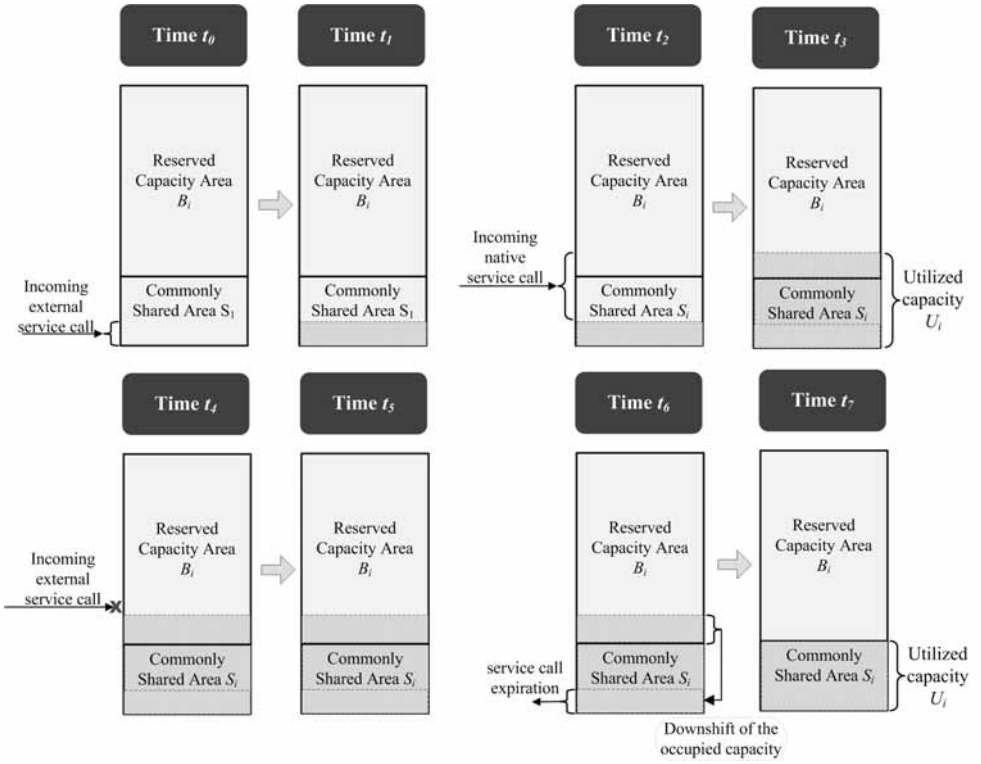

Figure 4.4: Capacity allocation examples assuming external and native service calls

The process of allocating capacity to incoming service calls starts from the S_i area where external and native calls can be accepted. When this area becomes fully occupied, then it continues by allocating capacity from the B_i area, however in this case only to native service calls. Consequently, if $R_{k,i}$ $(1 \leq k \leq K)$ is the capacity allocated to each of the K admitted calls to partition i, then the available capacity $C_{n,i}$ for a native service call is:

$$C_{n,i} = P_i - \sum_{k=1}^{K} R_{k,i} = P_i - U_{K,i} \qquad (4.3)$$

where $U_{K,i}$ is the total utilized capacity, in both areas, of the partition. Therefore, a native service call is able to utilize the total available capacity of the partition reduced only by $U_{K,i}$. On the other hand, the available capacity $C_{e,i}$ for an external service call is:

$$C_{e,i} = \begin{cases} S_i - \sum_{k=1}^{K} R_{k,i} = (1-b_i)\cdot P_i - U_{K,i} & , \text{if } S_i > U_{K,i} \\ 0 & , \text{if } S_i \leq U_{K,i} \end{cases} \qquad (4.4)$$

As a result, an external service call is able to utilize only the commonly shared area of the partition, which is also reduced by $U_{K,i}$. However, as $U_{K,i}$ is the total utilized capacity of the partition, it can become equal or greater than S_i. In this case, the available capacity $C_{e,i}$ for external service calls becomes zero and no external calls can be admitted to the partition.

An example of the capacity allocation process is shown in figure 4.4. Initially, at time t_0, the partition is empty and an external service call arrives requesting for capacity R_1. At this point, the available capacity for external calls is $C_{e,i}=S_i$. Therefore, assuming $R_1<C_{e,i}$, the call can be admitted to the S_i area of the partition (as indicated by the gray shading shown at time t_1). Subsequently, at time t_2 a native service call arrives requesting for capacity R_2. The available capacity for native calls is now $C_{n,i}=P_i-R_1$ and assuming that $R_2<C_{n,i}$ the call can also be admitted to the commnly shared area of the partition. However, if the remaining capacity at the S_i is not sufficient to accommodate the call, the latter can be extended to both areas of the partition as shown at time instance t_3. At time t_4 the incoming external call is blocked as $U_i>S_i$ and therefore there is no available capacity for external calls ($C_{e,i}=0$). Finally, at time instance t_6 a call, occupying capacity at the S_i area, departs. Then, as implied by equation (4.4), there is a downshift of the occupied capacity from the reserved area to the commonly shared area. Thus, as long as $U_i>S_i$ the commonly shared area is considered to be occupied and can not be utilized by external service calls.

4.3.2.4 Combined capacity

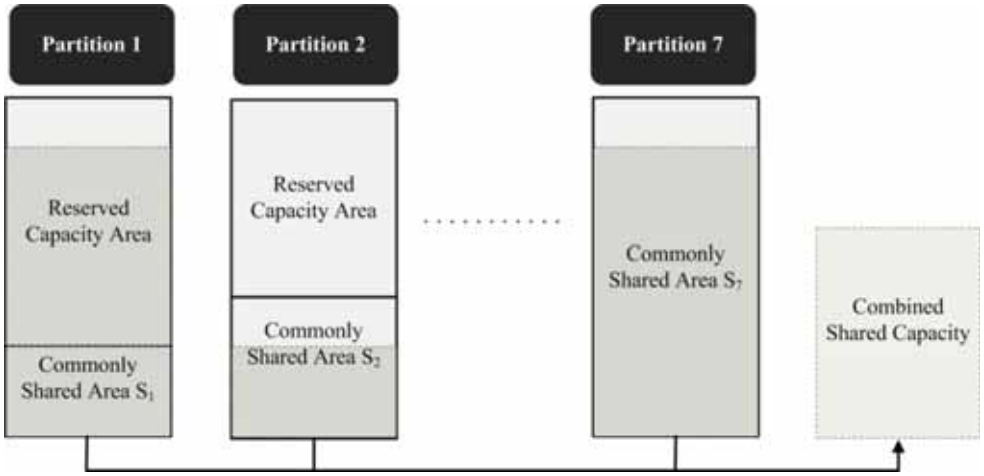

Figure 4.5: An incoming service call can be accepted by combining capacity from multiple partitions

In order to further increase the utilization of the available capacity, the notion of combined capacity C_C is also introduced. Thus, if the available capacity of a single partition is not adequate for the admission of a service call, then the system can accommodate the call by combining the available capacity from multiple partitions. The maximum combined capacity depends not only the occupation state of the partitions but also on the type of the incoming call. Therefore assuming an incoming service call which is native to partition l, ($1 \leq l \leq 7$), the maximum combined capacity it can utilize is:

$$\max\{C_C\} = C_{n,l} + \sum_{i=1, i\neq l}^{7} C_{e,i} \qquad (4.5)$$

Hence, a call is able to utilize the summation of the available capacity at its native partition ($C_{n,l}$) and the available parts of capacity ($C_{e,i}$) at the remaining partitions. Thus, the fragmentation of the system's capacity is eliminated, which in turn leads to optimal use of the available network resources. In figure 4.5, an example of combined capacity is shown assuming the first partition as native partition for the incoming call.

4.3.3 DSAC algorithm

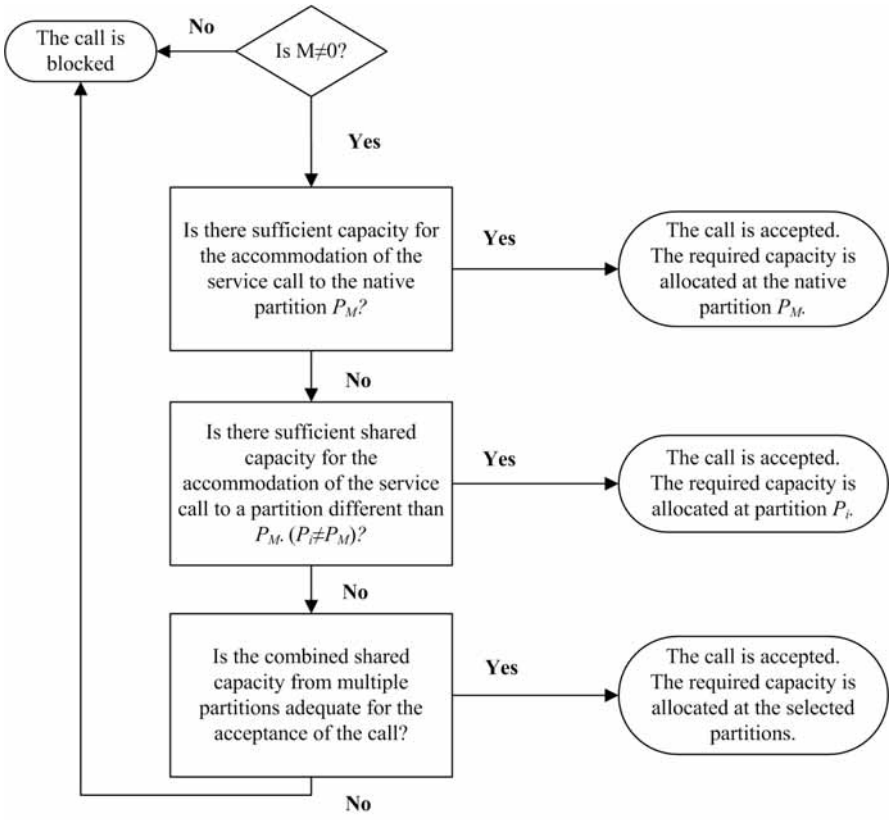

Figure 4.6: Flowchart of the Dynamic Service Admission Control (DSAC) scheme

On the previous subsections, the framework on which the proposed DSAC scheme operates was defined. In this subsection, the actual steps of the service admission control algorithm are described:

Upon the arrival of a service call $V_{i,j}$ that belongs to service class SC_i ($1 \leq i \leq 6$) and it is initiated from group j, ($1 \leq j \leq N$), DSAC has to check its partition mapping M. If M is set to zero (M=0), then the service call is blocked and the process stops. If $M \neq 0$, the algorithm initially checks if the service call can be accommodated in the partition that is mapped. If not, then the algorithm sequentially checks if the call can be accommodated to the commonly shared area of the other partitions. If again the call cannot be served, then the algorithm

checks if it can be accommodated by combining the shared capacity of multiple partitions. A flowchart of the Dynamic Service Admission Control scheme is shown in the figure 4.6.

4.3.4 Simulation environment setup

The performance of the proposed scheme is evaluated through event-driven simulation. Service calls are assumed to arrive according to a Poisson process, while their duration is exponentially distributed. All the parameters of the admission control framework as they are analyzed at the previous subsections are input data for the simulator. Therefore, depending on the studied scenario, two or more user groups can be defined each of them requesting calls from the available SCs. The backhaul capacity can be partitioned according to the composition of the offered traffic load, while a reservation factor *b* can also be defined for each of the partitions.

The evaluation of DSAC is performed in comparison with two other typical call admission control schemes, namely the Open Access Scheme (OAS) and Fixed Access Scheme with Common Pool (FAS-CP). The OAS scheme utilizes a single capacity partition, which contains all the available capacity. In this case, there isn't any capacity reservation for specific SCs and the call acceptance depends only on the total available capacity. On the contrary, FAS is a conventional capacity partitioning scheme, where a partition is defined for each service that requires QoS guarantees. The remaining capacity, obtained by subtracting the summation of the capacities of the partitions from the available backhaul capacity, is utilized as a common pool.

4.3.5 Performance evaluation results

4.3.5.1 Providing QoS differentiation

In order to study the QoS differentiation ability of the proposed DSAC scheme, a simple version of the enterprise scenario is assumed, where the users of the 1st CSG subgroup perform only voice calls (SC 3) through the femtocell. On the other hand, the users of the 2nd CSG group use the wireless/wireline network for video-conferencing calls (SC 1) and www browsing (SC 5). The probability of an incoming call to request for one of these services as well as the rate requirement and average duration of each kind of service is shown at table 4.3. Voice and video calls are required to experience a blocking probability of less than 0.01 (1%), while the www service calls to be treated in a best-effort manner, without QoS guarantees.

Therefore, for the FAS-CP and DSAC schemes, two capacity partitions are created. These partitions correspond to voice and video calls and reserve sufficient capacity so as to provide the required blocking probability at peak traffic load. The remaining capacity is utilized as a common pool partition for all the incoming calls including the www users. Furthermore, for

the DSAC scheme, a reservation factor *b* of 0.4 (40% of the initial partition capacity) is set for each of the first two partitions.

Table 4.3: Summary of simulation parameters

Type of service	Rate (kbps)	Percentage	Initiation Group
Voice	25	40%	1st CSG subgroup
video-conferencing	384	40%	2nd CSG subgroup
www browsing	120	20%	2nd CSG subgroup

In figures 4.7 and 4.8, the blocking probability of voice calls and video-conferencing calls for increasing traffic load is shown respectively. As expected, due to the reserved capacity partitions, the blocking probability of both services for the DSAC and FAS-CP schemes is less than 0.01 at all traffic loads. In contrast, the OAS scheme accommodates the incoming calls at a common partition without reserving any capacity for voice or video-conferencing calls. However, due to their low rate requirement the voice calls are experiencing low blocking probability, which exceeds 0.01 only at high traffic loads. Conversely, the video-conferencing calls due to their high rate requirement are experiencing high blocking probability which reaches 0.2 (20%) at peak traffic load. Therefore, OAS is not able to provide the required QoS for either the video-conferencing or voice calls.

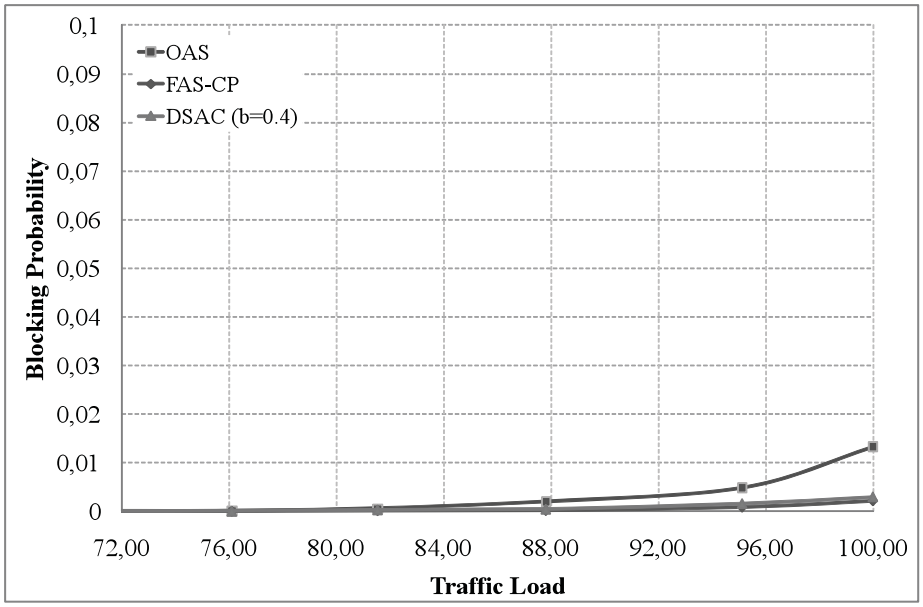

Figure 4.7: Blocking probability of voice calls

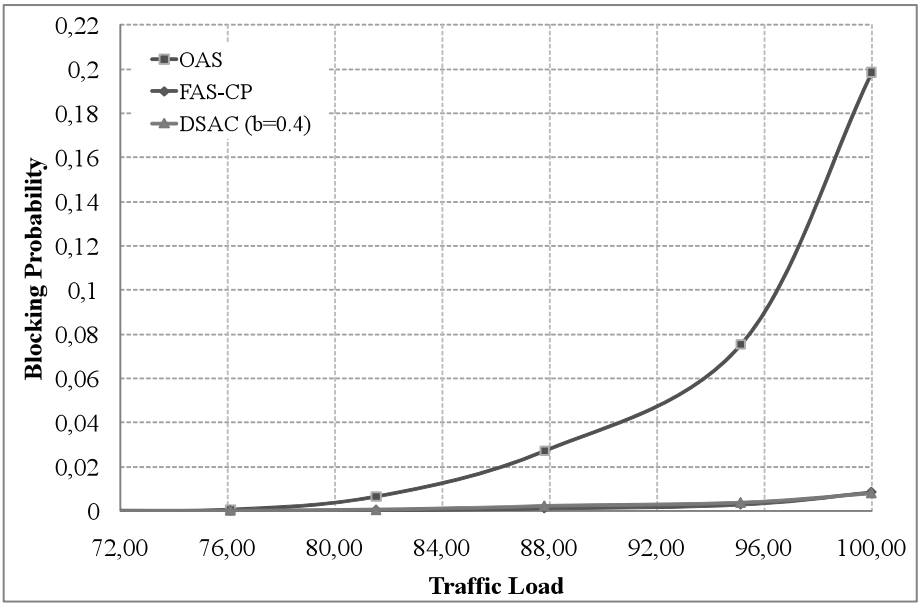

Figure 4.8: Blocking probability of video-conferencing calls

In figure 4.9, the blocking probability of www service calls is shown. Due to the formation of capacity partitions, the available capacity for the www users is considerably reduced when the DSAC and FAS-CP are employed in comparison to the available capacity when OAS is employed. Consequently, it is expected the respective blocking probability of DSAC and FAS-CP to be higher than that of the OAS scheme as it can be verified at figure 4.9. Nevertheless, DSAC, due to its ability to dynamically utilize the available capacity at the reserved partitions, outperforms FAS-CP at all traffic loads.

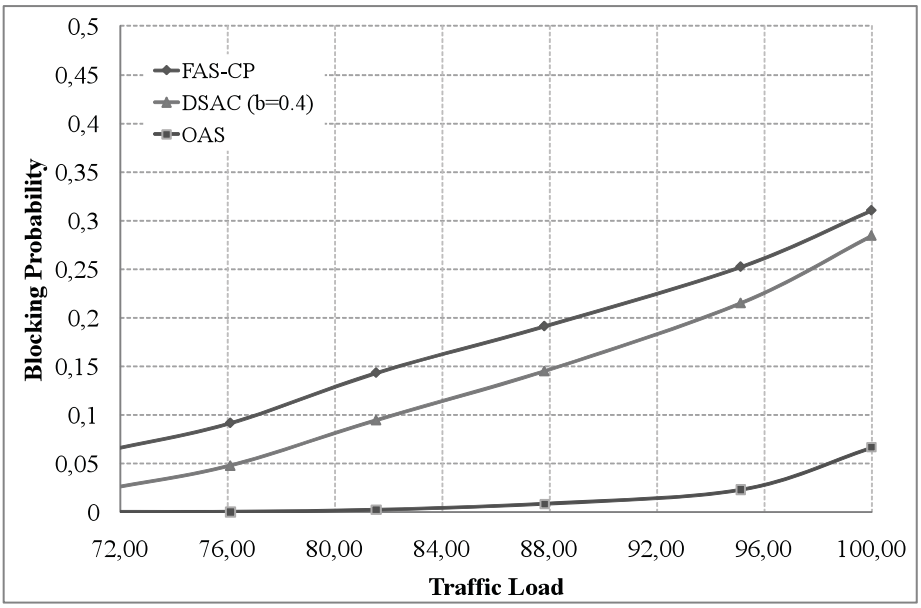

Figure 4.9: Blocking probability of www calls

Conclusively, while both FAS-CP and DSAC are able to provide the required QoS to voice and video-conferencing calls, the latter is also able to provide better QoS to the www users while simultaneously offering higher capacity utilization. This can be confirmed at figure 4.10, where the total capacity utilization for the three admission control schemes is shown. DSAC has better performance than FAS-CP at all traffic loads while, as expected, the OAS scheme, without reserving any capacity for QoS provisioning, outperforms both DSAC and FAS-CP.

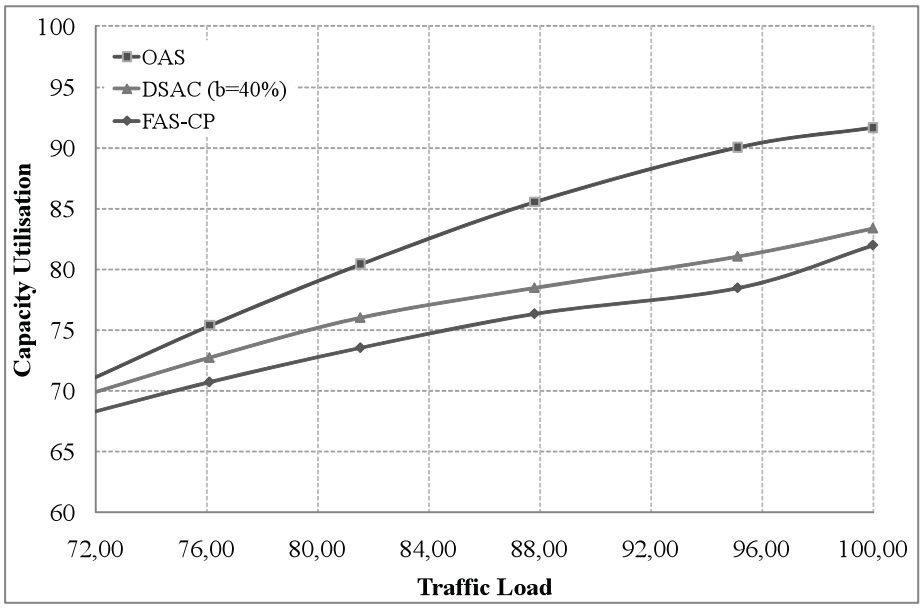

Figure 4.10: System capacity utilization

4.3.5.2 Confrontation of the short-term variations of the traffic load composition

We have showed so far that only the DSAC and FAS-CP schemes are able to provide QoS guarantees to the incoming calls and at the same time it has been shown that DSAC performs better than FAS-CP under normal traffic conditions. However, the main advantage of DSAC, against conventional partitioning schemes such as FAS-CP, is its ability to confront any short-term variations of the traffic load composition that may occur. Therefore, at the traffic scenario of the previous subsection (4.3.5.1), we assume that all the users of the 2nd CSG, temporarily, do not perform any video calls and they use the wireless/wireline network only for web browsing. The simulation parameters are summarized at table 4.4.

Table 4.4: Summary of simulation parameters

Type of service	Rate (kbps)	Percentage	Initiation Group
Voice	25	40%	1st CSG subgroup
video-conferencing	384	0%	2nd CSG subgroup
www browsing	120	60%	2nd CSG subgroup

As there is no “native” traffic load for the second partition, which was initially formed for the video-conferencing calls, the respective capacity is unutilized when the FAS-CP scheme is employed. In contrast, DSAC by using a reservation factor *b* of 0.4 allows the 60% of the second partition to be utilized reducing thus the blocking probability of the www users and increasing the overall system’s capacity utilization. This conclusion can be verified at figures 4.11 and 4.12 respectively. As one may see in figure 4.11, the blocking probability of www users ranges between 0.01 and 0.25 when the DSAC scheme is employed while for the FAS-CP the blocking probability ranges between 0.35 and 0.55. Furthermore, as shown in figure 4.12, the system capacity utilization under the DSAC scheme ranges between approximately 68% and 77% while for the FAS-CP scheme does not exceed 48%.

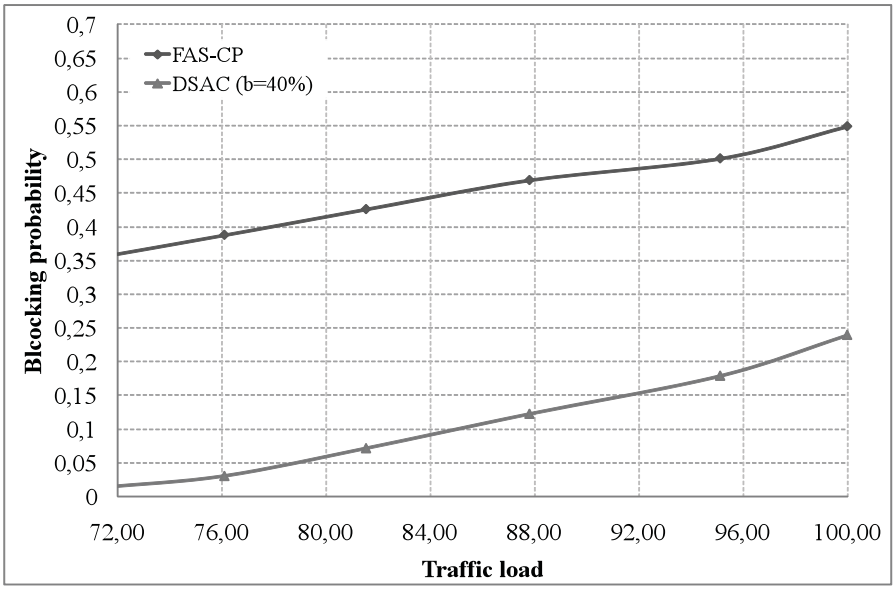

Figure 4.11: Blocking probability of www calls

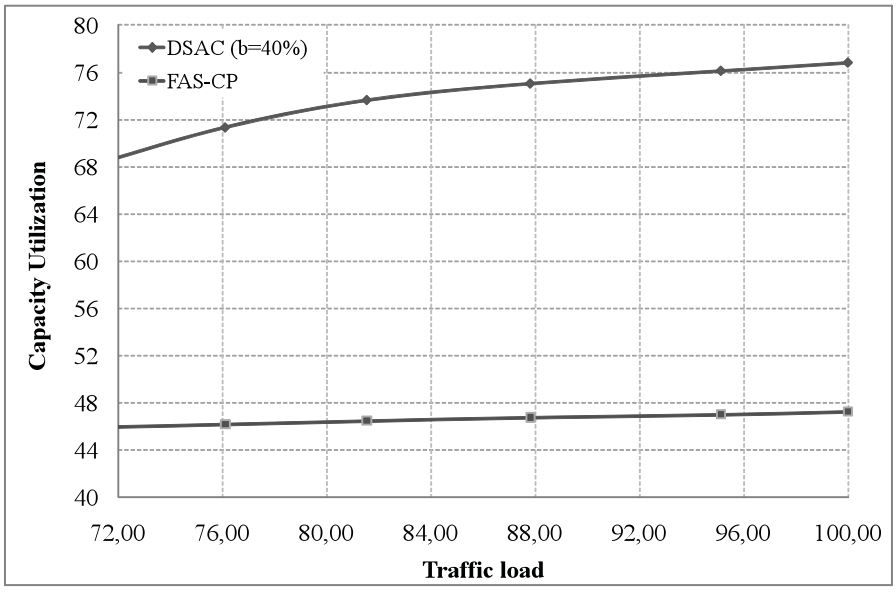

Figure 4.12: System capacity utilization

Conclusively, DSAC is able to effectively confront short-term variations of the traffic load composition offering significantly better performance in terms of blocking probability and system capacity utilization compared to conventional capacity partitioning schemes such as FAS-CP.

4.3.5.3 Performance of DSAC for varying value of the reservation factor (b)

By altering the value *b* of a partition's reservation factor, we can vary its usage from a completely shared partition, which equally accepts external and native service calls (*b*=0) to a completely shielded partition, which does not accept any external service calls (*b*=1). Thus, the performance of DSAC can significantly vary according to the value of *b.*

In order to demonstrate the ability of DSAC to vary its performance, the same traffic scenario as in subsection 4.3.5.1 is assumed, while we consecutively set *b* equal to 0.1 (10%), 0.2 (20%) and 0.3 (30%) for all the reserved partitions. At figures 4.13, 4.14 and 4.15, the blocking probabilities for voice, video-conferencing and www service calls are shown respectively. The reduction of the reservation factor *b* increases the commonly shared area of the reserved partitions and that leads to higher blocking probability for the voice and video-conferencing calls and lower blocking probability for the www service calls. Therefore, as *b* decreases, one may observe the plot in figures 4.13 and 4.14 to be progressively upshifted, while at figure 4.15 the plot is downshifted. Similar behaviour can be observed at figure 4.16 where the overall system capacity utilization is shown. As the reservation factor is decreased, less capacity is reserved for QoS provisioning and therefore the capacity utilization is increased.

Conclusively, the value of *b* is highly related to the assumed network deployment scenario. For example, at the home user scenario (described at subsection 4.2.1.1), the value of *b* will generally be high in order to shield the services provided to the CSG users. Furthermore, for the enterprise scenario (described at section 4.2.1.2), high values of *b* can be applied for the services offered to the employees, while moderate values of *b* can be applied for the services offered to guest users. Finally, in the case of the public access scenario (described at section 4.2.1.3), a low value of *b* would be beneficial in order to serve as many users as possible and better utilize the available capacity. However, a moderate value of *b* can also be set for users who want to make use of, resource consuming, personalized services offered by the network. An outline of the reservation factor *b* values for each of the three discussed network deployment scenarios is shown at table 4.5.

Table 4.5: Scenario depended value of the reservation factor (*b*)

Scenario	Reservation Factor Value
Home User	High
Enterprise	High to Moderate
Public Access	Moderate to Low

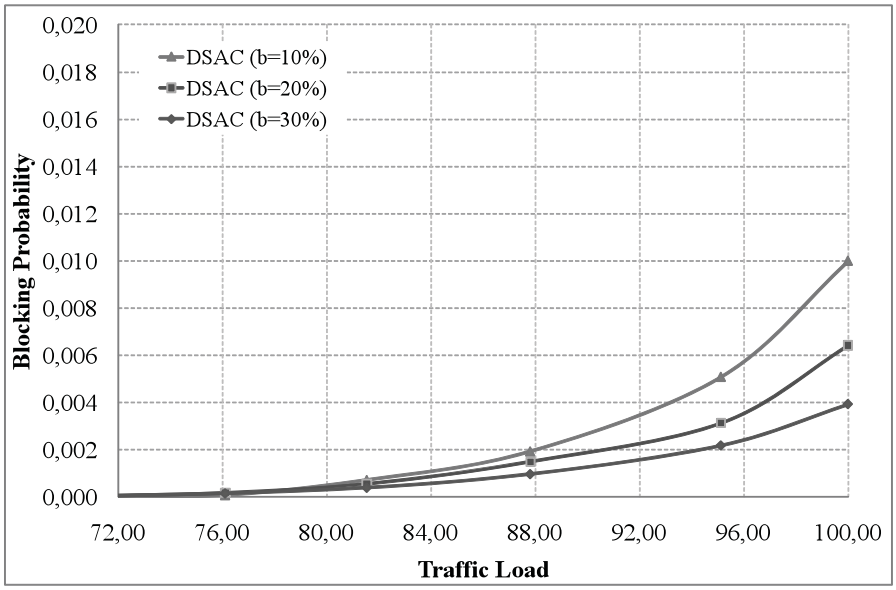

Figure 4.13: Blocking probability of voice calls

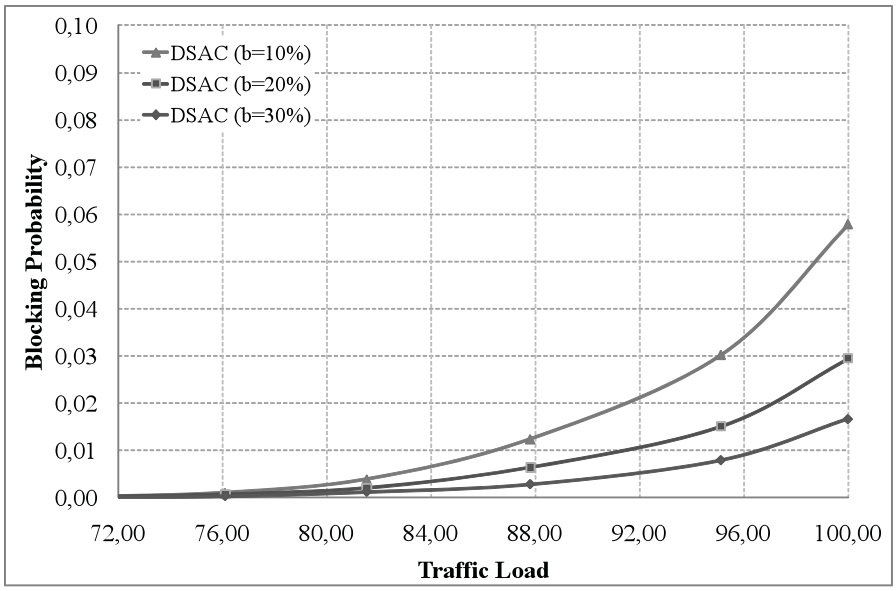

Figure 4.14: Blocking probability of video-conferencing calls

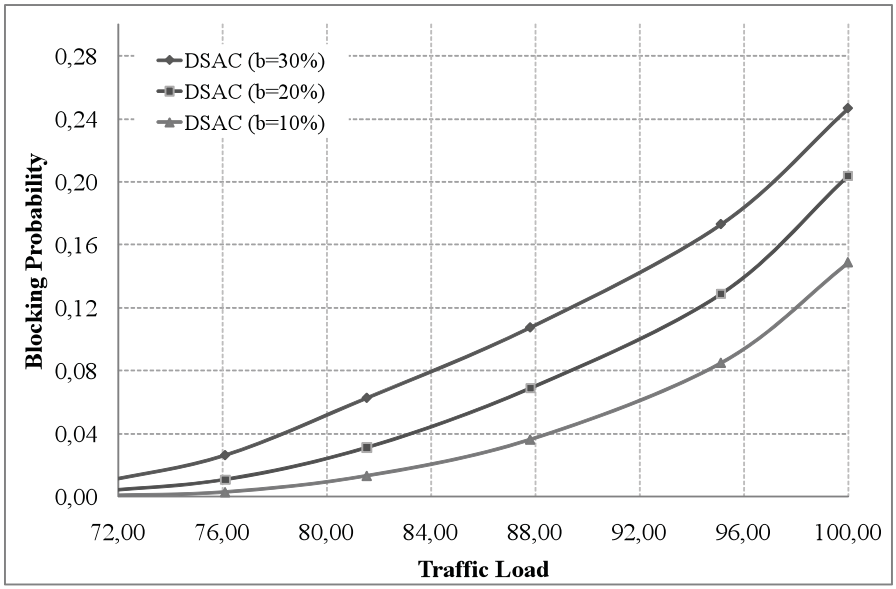

Figure 4.15: Blocking probability of www service calls

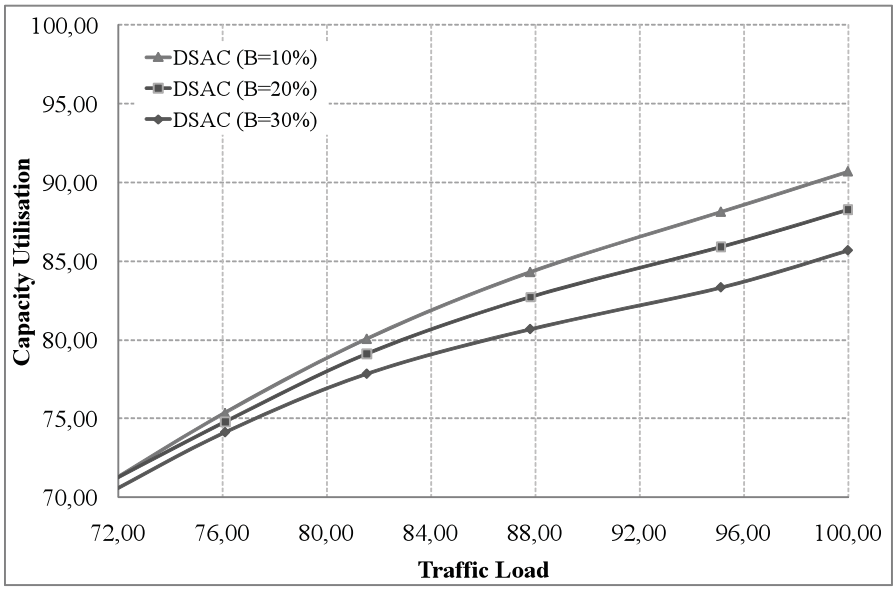

Figure 4.16: System capacity utilization

4.4 Proposed Integrated Services Admission Control (ISAC) Scheme

The system model depicted in figure 3.6 is assumed for this section. Here, the focus is on the RRM integration layer of CA-FEI framework (while in chapter 3, relay module's operations are described regarding interference management issues) as this was described in section 3.3 (cf. figure 3.7). Proposed ISAC scheme cooperates with context information acquisition (CIAM) and relay modules and hence one has to be aware of all modules' operations in order to understand the functionalities of proposed CA-FEI framework as a whole.

ISAC scheme is composed of two robust, efficient and low complexity modules, which can be easily implemented and directly applicable in real-market femtocell deployment scenarios. The first module is a service admission control (SAC) module, which is based on the

effectiveness of well-known capacity partitioning rationale, while it is complemented with a novel periodic partition adjustment (PPA) process that addresses any backhaul capacity fluctuations as well as variations of traffic load composition. Although the main concern is to preserve the QoS of delay sensitive services originating from the femtocell, we also took into account that similar services may originate from other network interfaces and, therefore, they have to be treated equally. Hence, SAC was designed to operate over a single virtual network interface, applying thus the same admission policies to the total incoming traffic load. Subsequently, the second module, called capacity management module (CAM), interprets the SAC admission decisions to the actual capacity distribution among the real physical network interfaces (figure 4.17).

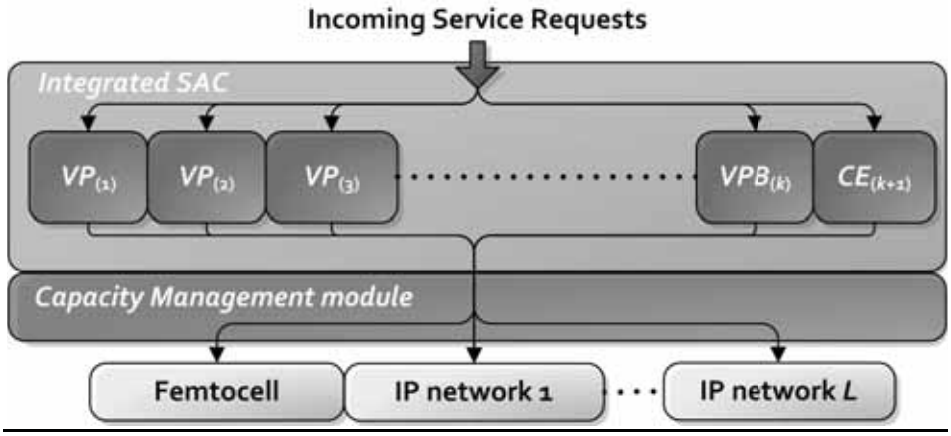

Figure 4.17: The ISAC scheme of CA-FEI framework

4.4.1 ISAC's interactions with other CA-FEI framework's modules

As shown in figure 3.7, the RRM integration layer (i.e. ISAC scheme) has interfaces with the: a) various network access technologies (i.e. or else a single virtual network interface) in order to be aware of the radio resources requests, b) IP router in order to be aware of the available wired backhaul capacity, and c) context information acquisition module (CIAM). CIAM acquires low-level data from the IP router and the network interfaces, and transforms them to the proper form (i.e. high-level context information) in order to be utilized as input for the ISAC modules as well as for the relay module. For example, in cases where CIAM identifies high wired backhaul capacity utilization, it can trigger the relay module to route delay sensitive traffic or traffic incurred by macro UEs via the wireless backhaul link (see also section 3.3). On the contrary, when the relay module informs CIAM that the wireless backhaul link is congested or the femtocell's radio environment experiences high interference levels, CIAM can inform the RRM integration layer in order the various partitions' capacities to be restructured in ways that efficient QoS can be achieved for all types of services. CIAM has two main functionalities regarding context acquisition, which are: a) initial setup phase, and b) monitoring/gathering of context.

Regarding the initial setup and during the deployment phase, the user (i.e. home user or ICT technician) will be able to set, through a GUI, the initial operating parameters. Such parameters are the number of users that is expected to be served by the femtocell as well as their preferences regarding the services they usually use and the way they use them. As this can be a difficult task for the average home user, he will be offered, through the setup GUI, a set of typical values to select from. More advanced setup procedures may even autonomously select one of these sets through a question–answer process interacting with the user, aiming to discover his preferences. Furthermore, an "advanced user mode" will allow more experienced users or ICT specialists to finely tune the overall process. The result of this initial setup phase will serve as a basis on which the CIAM will build its future outcomes. In cases, where the home user has to configure the initial setup phase, he/she can assign priorities to specific UEs, which frequently utilize the femtocell for acquiring access to mobile services (e.g., the home user can define his/her CSG and determine the context in which a nearby macro UE can connect to the femtocell without degrading the quality of experience (QoE) of CSG's users). Another priority assignment procedure can take place for different service classes (e.g., the home user can define which type of services will have the highest priority and thus no QoS degradation is acceptable even in high backhaul capacity utilization situations). Given the fact that a mentality shift occurs in cellular operators' industry towards giving several motivations to femtocell owners in order their FAPs to serve more macro UEs and thus offload mobile traffic from cellular operators' infrastructure, the initial setup phase can become very important for the femtocell owner to experience considerable bill discounts. In cases, where ICT specialists configure the initial setup process (e.g., public access and enterprise scenarios analyzed in 4.2.1), the whole procedure can become even more complicated as hybrid or open access femtocells are assumed and thus service classes' and user groups' priorities should be configured taking into consideration the overall HetNet environment, too.

Regarding context monitoring and gathering and during operation phase, CIAM keeps track of the statistical properties of the incoming traffic load as well as of the backhaul capacity in order to be able to adjust, if needed, the values initially selected by the user. For example, capacity adjustments and the various partitions' restructure can happen in cases of overload-state situations or in cases when context-aware prioritization based on the type of services that the femto-relay is currently supporting, occur. More specifically, if wired backhaul experiences intolerable delays (e.g., in overload-state situations or when delay-sensitive services cannot effectively be delivered), then the wireless backhaul is selected, as the required QoS levels can be more easily maintained, because of the fact that the data "remains" within one cellular operator's network.

The first sub-functionality of monitoring is the sensing of context, meaning both the detection and reception of data from sensors without introducing additional traffic overhead. Efficient

scanning is a matter of vital importance since real-time limitations have to be satisfied and QoE objectives should be maintained within acceptable levels [23]. The key objective is the dynamic adjustment of the defined set of parameters by adopting event-based and on-demand monitoring policies [88]. One example of monitoring is the sensing of the FAP's radio environment in order to clarify if interference levels are below some predefined and desirable thresholds (see details in chapter 3). If this is the case, then the relay module can inform CIAM about the wireless backhaul link alternative and consequently CIAM will update the wired backhaul capacity partitions (i.e. PPA process), so that the latter can be restructured in favor of the QoS/QoE metrics for MTs/UEs. Another monitoring/gathering example is when CIAM monitors ISAC modules. More specifically, when a significant traffic load variation occurs or if the backhaul capacity utilization is relatively low or specific delay constraints for real-time services have to be met, CIAM decides the extent of wired/wireless backhaul usage based on the context information that receives from the relay module, too. Although the RRM process can dynamically be adapted to small variations of the traffic load distribution through the PPA process, the outcome of CIAM is essential for the efficiency of the context aware resource management operation under traffic load variations taking place in larger timeframes. Equally important is the outcome of CIAM for the activation or deactivation of the relay module. More specifically, when a FAP's utilization is relatively low at specific time intervals (cf. available femtocell case depicted in figure 3.4), CIAM triggers the relay module's operation and thus the femtocell can act as a relay allowing neighboring macro MTs/UEs to transmit in a two-hop fashion.

4.4.2 Backhaul capacity partitioning

In order to perform the ISAC operation, the definition of common service classes is required for all the underlying networks. Thus, services with similar characteristics, regardless of their network origin, are classified to the same Integrated Service Group (ISG). The number of ISGs should be at least four, corresponding to conversational, streaming, interactive and background service classes as determined by 3GPP [181]. However, each ISG can be further partitioned, if a more precise categorization is required. In the general case, corresponding to a total of k ISGs, the backhaul capacity is also divided to k primary virtual partitions (VPs) and each partition S_j, $1 \leq j \leq k$ is associated to the j^{th} ISG_j. An additional (k+1) partition, namely the capacity exchange (CE) partition, is also defined as required by the PPA process (cf. subsection 4.4.2.1). A prerequisite for ISAC scheme is that a method for indicating the QoS requirements of each traffic flow, such as DiffServ [220], should be employed at all coexisting networks. However, if a flow does not need any QoS or does not indicate that it needs a specific QoS treatment, then it will be treated as a background service.

The aim of ISAC is to ensure that a service call with specific QoS requirements is admitted to the system only if there is adequate capacity to handle it. Consequently, the initial virtual capacity partitioning should be performed for the worst-case situation, where the traffic load intensity A_j offered by each ISG_j has its maximum value, while at the same time the backhaul capacity has its minimum value C_{min}. During this critical system state, the *VP* of the delay insensitive background services, denoted in the following as *VPB*, is restricted to its minimum size. Assuming that *VPB* is the k^{th} partition, then the partitioning problem is constrained to the first *k-1 VP*s.

Let a_j be the average rate requested by services of ISG_j, which we assume that arrive following a Poisson process with exponentially distributed holding times. Then, the initial size S_j^0 of partition VP_j can be defined as a multiple of a_j $\left(S_j^0 = m_j \times a_j, m_j \in \mathbb{Z}^+\right)$ and for a given value of m_j the blocking probability $P_{b,j}(A_j,m_j)$ for the members of the ISG_j class can be computed by the Erlang B formula. If G_j is the Grade of Service (*GoS*) required for ISG_j services, then the overall partitioning problem can be described as the definition of the vector $m_j = (m_1, m_2, \ldots m_{k-1})$, which satisfies inequalities (4.6) and (4.7):

$$P_{b,j}\left(A_j,m_j\right) \le G_j, \quad \forall j \in \left[1,\left(k-1\right)\right] \tag{4.6}$$

$$\sum_{j=1}^{k-1}\left(S_j^0\right) \le C_{\min} \tag{4.7}$$

Although, in the general case, the above set of equations has multiple possible solutions, the initial size of the partitions can be defined through a straightforward process that ensures a reliable operation in the worst-case scenario.

(a) At first, the size of each of the $k-1$ partitions has to be defined in such a way that the required upper bound G_j of the blocking probability of the ISG_j calls to be guaranteed. Therefore, as the initial size of the partitions should correspond to the maximum acceptable blocking probability, each S_j^0 can be calculated directly from equation (4.6) for the case of equality (i.e., $P_{b,j}(A_j, m_j) = G_j, \forall\, j \in [1, (k-1)]$).

(b) Subsequently, it has to be ensured that the summation of the calculated capacities S_j^0 of the partitions meets the limitation of equation (4.7). If not, then the network administrator has to select and prioritize those *ISGs* for which the target blocking probability can be further increased. The administrator has to take into account both the nature of the network and the requirements of the users so that his decision to be as tailored as possible to the specific network environment. In case the femtocell is installed by a home user, then he should be guided by a properly designed setup process in order to make the right selections based on his preferences (cf. section 4.4.1). Given the prioritized list of the selected *ISGs*, the order of the

list is iteratively followed, and at each iteration the size of each partition S_j^0 is reduced by a_j until equation (4.7) is valid.

(c) Finally, if a part C_L of the $C_{\min}$ capacity is left unallocated after the initial partitioning process, then it is shared among the j partitions proportionally to their respective traffic load intensity A_j. Thus, each partition is extended by a fraction of C_L, which is normalized to a_j. The initial size S_j^0 of each partition VP_j can be defined as:

$$S_j^0 = \left(m_j \cdot a_j\right) + \left(l_j \cdot a_j\right) = \left(m_j + l_j\right) \cdot a_j \qquad (4.8)$$

where l_j is an integer, which is calculated by applying the floor function as follows:

$$l_j = \left(A_j / \sum_{i=1}^{k} \left(A_i\right) \right) \cdot C_L \cdot \frac{1}{a_j} \qquad (4.9)$$

The additional capacity is calculated consecutively starting from the partition that is expected to handle the higher traffic load, while ties are randomly resolved.

4.4.2.1 Periodic partition adjustment (PPA) process

While the initial capacity partitioning guarantees a reliable operation for the worst-case scenario, at the same time it could lead to capacity underutilization under normal traffic conditions. Therefore, the PPA process is employed in order to sequentially transfer unutilized capacity from the $k-1$ primary *VPs* to the *CE* partition and vice versa. Consequently, the *VPs*' size is not fixed, but instead it is adjusted in response to the variations of the traffic load composition.

Starting from VP_1 towards VP_{k-1}, the PPA process, at time intervals of Δt, sequentially adjusts the size of all *VP*s. Thus, the size $S_j(t)$ of VP_j partition is adjusted by a term $w_j(t)$:

$$S_j\left(t\right) = \min\left\{\left[S_j\left(t-\Delta t\right) + w_j\left(t\right)\right], S_j^0\right\} \qquad (4.10)$$

Therefore, as implied by equation (4.10), the upper bound of $S_j(t)$ is set to S_j^0 (cf. equation 4.8). The adjusting function $w_j(t)$ is equal to a negative or positive step of a_j depending on the blocking rate B_j that the ISG_j service calls have experienced, averaged over a time interval of d seconds, and the current occupancy O_j of the partition, namely:

$$w_j\left(t\right) = \begin{cases} -a_j \text{ if } \left[B_j\left(t-d\right) < G_j\right] \text{ and } \left[O_j\left(t\right) \le S_j\left(t\right) - g_j \cdot a_j\right] \\ 0 \text{ if } \left[B_j\left(t-d\right) < G_j\right] \text{ and } \left[O_j\left(t\right) > S_j\left(t\right) - g_j \cdot a_j\right] \\ a_j \text{ if } \left[B_j\left(t-d\right) \ge G_j\right] \end{cases} \qquad (4.11)$$

The higher the value of d is, the more averaged over time B_j becomes. The guarding factor g_j is a positive integer ($1 \le g_j \le m_j$), which guarantees that the size of the partition can be decreased only if it has at least $g_j \cdot a_j$ available capacity for the future incoming ISG_j service

calls. Furthermore, the size of a *VP* can be increased by a_j only if a corresponding decrease of the *CE* partition by $-a_j$, according to its occupancy, can be performed. If not, the capacity adjustment of the *VP* is scheduled and performed asynchronously, when conditions permit together with the inverse adjustment of the *CE*.

The total capacity adjustment for the *CE* partition is the negative of the summation of all the $w_j(t)$ adjustments calculated for the *VP*s. Hence, the required capacity $S_{CE}^{R}(t)$ for the *CE* partition is:

$$S_{\mathrm{CE}}^{R}(t) = S_{\mathrm{CE}}^{R}(t-\Delta t) + \sum_{j=1}^{k=1}\left(-w_j(t)\right) \tag{4.12}$$

However, the CE partition is not only used as a CE point between the *VP*s, but also acts as a common pool for all *ISG*s. As a result, the actual capacity $S_{\mathrm{CE}}(t)$ of the *CE* partition cannot become less than the currently occupied capacity $O_{\mathrm{CE}}(t)$:

$$S_{\mathrm{CE}}(t) = \max\left\{S_{\mathrm{CE}}^{R}(t), O_{\mathrm{CE}}(t)\right\} \tag{4.13}$$

As previously mentioned, the adjustment steps, which cannot be performed due to the restriction of equation (4.13), are scheduled and performed asynchronously.

Based on the previous steps, the partitioning process up to this point is dimensioned on the minimum backhaul capacity $C_{\min}$. The difference between $C_{\min}$ and the current value of the backhaul capacity $C(t)$ forms the *VP k* for the background services:

$$S_{\mathrm{VPB}}(t) = C(t) - C_{\min}(t) \tag{4.14}$$

Obviously, the background services can be accepted either at the *VPB* partition or at the *CE* partition.

4.4.3 ISAC algorithm

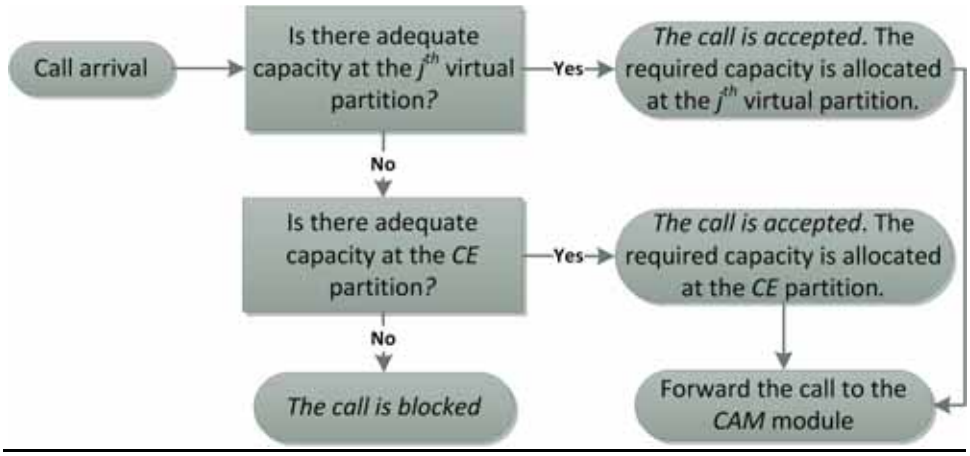

Figure 4.18: Flowchart of the ISAC algorithm

As long as the virtual capacity partitions are defined, the admission decision process, as shown at figure 4.18, is straightforward. An incoming service call of ISG_j is accepted only if: a) there is an adequate free capacity at the respective *j*th virtual partition or else b) if there is an adequate free capacity at the *CE* partition. Thus, the ISAC scheme ensures that: a) all service requests are handled equally irrespective of the employed network interface and that b) there is always adequate backhaul capacity to handle all the admitted service calls with their respective QoS requirements.

Consequently, ISAC extends the QoS-aware admission control of femtocells to all the underlying IP networks following the FI paradigm discussed at [221]. Furthermore, ISAC does not require the ISP to exclusively provide any kind of QoS guarantees. On the contrary, the main admission policy of ISAC is to limit the acceptance of QoS demanding services to what the ISP is proven to be able to provide. After a service call is accepted and has been allocated capacity at one of the *VPs,* it is then forwarded to the CAM.

CAM is responsible to route an admitted service call to the appropriate serving network and to adjust the backhaul capacity distribution among all the underlying networks. Specifically, assuming a total of L underlying networks (see figure 4.17) each one having N_y $(1 \le y \le L)$ ongoing service calls of average requested rate a_i $(1 \le i \le N_y)$, the capacity $U_y(t)$ allocated to each network is defined as follows:

$$U_y(t) = \left(\sum_{i=1}^{N_y} a_i \Big/ \sum_{x=1}^{L} \sum_{i=1}^{N_x} a_i \right) \cdot C(t) \qquad (4.15)$$

Hence, each admitted service call modifies the fraction of the current backhaul capacity, which is allocated to each serving network. Consequently, while the SAC module is based on the minimum value of the backhaul capacity in order to be able to guarantee a minimum QoS for all admitted service calls, CAM module distributes the total available backhaul capacity among the serving networks.

In other words, although a VP may be empty and reserved for a specific ISG, its respective physical capacity does not remain unused but instead can always be utilized by the underlying packet schedulers, as implied by Equation (4.15). Therefore, it becomes clear that designing CA-FEI framework as a two-tier scheme makes it possible to have, at the same time, a virtual (logical) capacity distribution for the admission control process and an actual (physical) capacity distribution for the serving networks. As a result, the proposed CA-FEI is more flexible compared to typical one-tier frameworks that utilize actual capacity partitions and thus it is able to achieve higher utilization of the backhaul capacity. Furthermore, it can also be concluded from equation (4.15) that if two networks have to serve the same traffic load, in terms of volume and distribution among ISGs, these networks will also be allocated the same

bandwidth. Consequently, the bandwidth distribution among the serving networks can be considered as fair because it is directly proportional to their ongoing traffic load.

4.4.4 Performance evaluation results

In this section, the ability of ISAC scheme to continuously adjust to variations of the traffic load distribution is studied. Hybrid Partitioning Schemes (HPS) and Fixed Partitioning Schemes (FPS) [203] can be considered as a base of reference for comparison because their operation can be approximated, up to some point, with the currently available in the market router and femtocell combinations. However, HPS offer better capacity utilization and are more immune against traffic variations compared to FPS and are thus better competitors for the proposed ISAC scheme. In our case, the employed HPS has a fixed partition, defined for each service, which requires specific QoS guarantees, while the remaining capacity is utilized as common pool partition for all service calls. Essentially, HPS design rationale is similar with DSAC scheme proposed in the previous section. Hence, it is shown how ISAC operation can outperform DSAC due to the employed novel PPA function of ISAC.

A simulation scenario is assumed, where a femtocell coexists with a wireline IP network sharing a 100 Mbps fast ethernet backhaul line. The period Δt of PPA is set to 100 ms, the guarding factor is uniformly set to 10 and the averaging time interval d is extended up to the duration of the simulation (i.e. the blocking rate B is recalculated each time PPA is executed). Both ISAC and HPS schemes arrange their capacity partitions expecting at the worst case a 35% reduction of the backhaul capacity as well as a traffic load distribution consisting of only delay sensitive services (40% Voice, 30% Video-telephony, and 30% Streaming video) all having a target GoS of 1%. In order to analyze the effect of PPA function, we further assume that the traffic load distribution presents an unpredictable short-term variation. Thus, the percentage of video-telephony calls is increased to 50%, while the percentage of streaming video calls is decreased to 10%. The following simulation results correspond to a confidence interval of less than 5% with a respective confidence level of 95%.

Due to the assumed worst case scenario and the strict required GoS, the capacity left for the common pool partition is significantly reduced for both HPS and ISAC schemes. Thus, the HPS scheme cannot cope with the unexpected variation of the traffic load. As a result, the blocking probability of video-telephony calls ranges approximately between 52 and 58% when the backhaul capacity is highly utilized (as shown in figure 4.19). At the same time, the reduced streaming video traffic makes use of a capacity partition that is initially designed to handle a significantly higher traffic volume. Consequently, streaming video calls have a zero blocking probability.

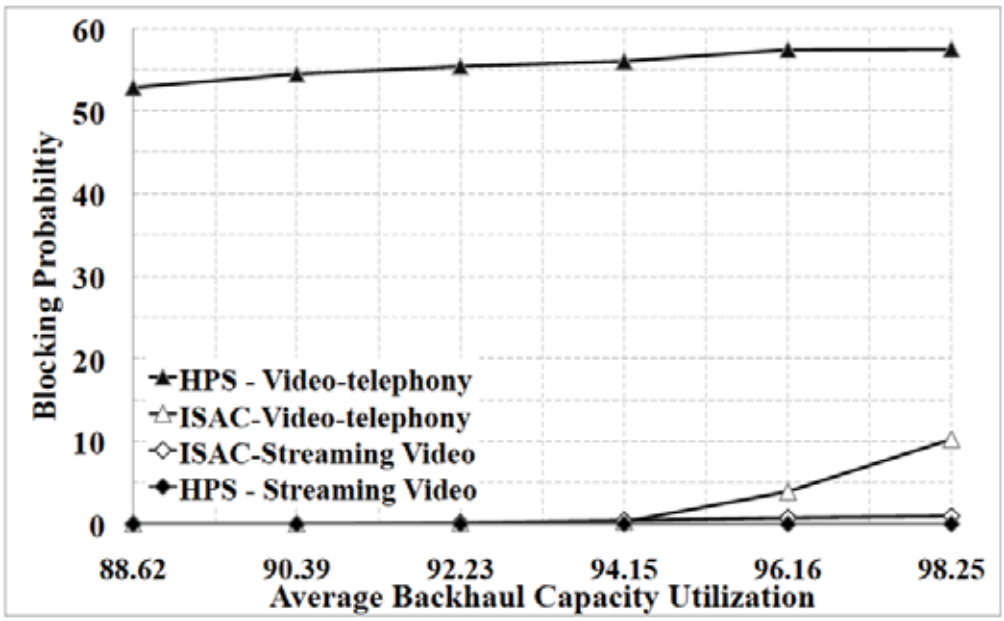

Figure 4.19: Blocking probability (HPS vs. ISAC scheme)

On the other hand, when ISAC scheme is employed, video-telephony calls are able to utilize the capacity, which remains unused at the partition of streaming video calls. Consequently, the blocking probability of video-telephony is significantly reduced compared to HPS and remains below the required GoS of 1% for traffic loads up to 94%. At the same time, due to the employed guarding factor, the blocking probability of the streaming video calls remains also below 1% at all traffic loads, even though their partition is in this case fully utilized. The efficiency of the ISAC scheme can also be verified at table 4.6, where it is shown that ISAC offers increased capacity utilization compared to the HPS scheme.

Table 4.6: Overall system capacity utilization

Traffic load	88.62	90.39	92.23	94.15	96.16	98.25
HPS	48.77	49.40	49.16	49.70	50.33	50.81
ISAC	84.06	85.17	87.12	87.52	88.12	85.65
% of increase	72.35	72.40	77.22	76.10	75.08	68.57

In conclusion, at high traffic loads, HPS (and consequently FPS, used in current femtocell deployments) are inefficient in terms of capacity utilization. Hence, considering the increased traffic volume expected for LTE systems, a novel framework has been proposed, which provides a suitable environment for the femtocell operation through effective and efficient management of the total incoming traffic and the common limited backhaul capacity.

4.5 Proposed Context Aware Backhaul Management (CABM) Scheme

The use of today's networks to connect machines/things to the Internet is still in its infancy, while a continuous increase in mobile data traffic generation is observed, fueled by the increasing number of sensors and the use of networks for Machine-to-Machine (M2M) communications. According to conservative estimations, cellular M2M connections are expected to overcome 200 million in 2013, while trillions of Machine Type Communication Devices (MTCDs) will exist by the end of this decade [28]. The integration of M2M communications' concepts to existing Human-to-Human (H2H)-oriented network infrastructures is still a challenge for standardization bodies such as 3GPP [222] and ETSI [223]. For example, some major requirements that the M2M communication paradigm introduces are [155], [224]: a) the huge number of MTCDs, which is much larger than cellular devices, b) MTC services often have different QoS requirements due to their specific features, c) traffic models generated from MTCDs differ from that generated by cellular devices, d) most of the traffic occurs in the uplink, e) MTCs involve innovative real-market applicability scenarios, and f) MTCDs often operate on restricted power, and thus the adoption of energy efficient techniques is necessary.

In this section, the focus is on context aware resource management issues for combined Human-to-Human (H2H) and Machine-to-Machine (M2M) traffic. The proposed Context-Aware Backhaul Management (CABM) scheme is applicable for M2M gateways, which can provide the interconnection between cellular and capillary networks both via wired and wireless backhaul alternatives (see figure 4.1). CABM scheme adopts efficient QoS provisioning procedures, by taking into consideration the diverse QoS requirements of novel MTC services. Simulation results show that the proposed MTC Request Admission Control (MTCR-AC) algorithm is resilient regarding QoS-related metrics in overload-state situations, while its inherent prediction-based decision-making procedures can offer significant performance gains, too.

4.5.1 Related work overview and problem formulation

As depicted in figure 4.1, the core architectural element of the assumed system model is the MTC Gateway. MTCG's operation has been already introduced by international standardization bodies such as ETSI [223] and 3GPP [225]. More specifically, there are several high-level descriptions and design requirements regarding MTCG's capability to coordinate the whole communication between MTCDs' and core network domains [28]. There are also clear implications about MTCG's capability to perform resource provisioning, congestion control and adaptation of MTC services to diverse contexts (i.e. service type, MTCD type, network status, etc). However, the above-mentioned technical specifications

don’t analyze in-depth the assessed high-level descriptions and thus leave their implementation to the research community.

There have also been efforts to differentiate MTCGs’ operation based on assumed real-market applicability scenarios/use cases. For example in [218], the scenario of residential M2M networks is addressed and the design requirements of a home MTCG are provided. In [155], a variety of architectural enhancements to LTE-A cellular networks with M2M communications is provided, while it is shown that the role of a MTCG is expected to be critical in public access and wide geographical range scenarios. EU co-funded EXALTED project [215] has also identified the critical role of a MTCG in various deployment settings, while 3GPP specifications themselves describe specific M2M network scenarios/use cases, which involve MTCGs [225].

Congestion control and radio resource management (RRM) challenges for MTC over LTE or other broadband networks have also been recently addressed in the international literature. Congestion/overload control issues at the cellular operator’s evolved packet core (EPC) network are presented in [226] [227] emphasizing on different congestion locations such as the LTE base station (eNB) and the EPC gateways (i.e. MME, S-GW and P-GW) but not on the MTCG. Novel RRM algorithms (i.e. scheduling, admission/rate control, etc) residing at the eNB have been proposed during the last couple of years, too. More specifically, works such as [228] and [229] deal with scheduling techniques for MTC over LTE, [155] presents various RRM schemes, [218] proposes a cross-layer joint admission and rate control scheme and [230] proposes a RRM scheme with QoS guarantees for MTC services. QoS provisioning and the need for novel techniques regarding QoS classes formation for combined H2H and M2M traffic is stressed in [224] and [231] but no specific implementation and proof-of-concept results have been conducted so far.

The proposed Context-Aware Backhaul Management (CABM) scheme is illustrated in figure 4.20. As previously stated, CABM is applicable for a MTCG in a network architecture such as the one depicted in figure 4.1. MTCG is assumed to have all well-known heterogeneous wireless access network technologies’ interfaces such as cellular (e.g. 2G, 3G, 4G), WiFi (e.g. 802.11x), WPAN/WBAN (e.g. 802.15.x, ZigBee) and wired (e.g. ethernet) interfaces. CABM consists of two main modules, which are: a) the Context Management (CM) module, and b) the Backhaul Management (BM) module. Once a MTC or conventional H2H service request arrives at the MTCG, the CM analyzes the context (i.e. service type, MTCD type, network status, etc) and informs the BM about the QoS class that the MTC service belongs to. The BM monitors the current available backhaul (either wired or wireless) capacity and decides whether the incoming service request can be accepted or rejected and under which terms the resource allocation procedure will take place. After the decision has been made, BM informs CM correspondingly and the latter makes the appropriate actions based on updated context

such as sending signaling messages to neighboring MTCGs, single MTCDs and/or MTCD groups. CM module adopts all context-aware mobile and wireless networking principles described in [23]. For example, by having prediction and learning capabilities, it can make more intelligent decisions subsequently enhancing the performance of BM, too (see more in table 4.8).

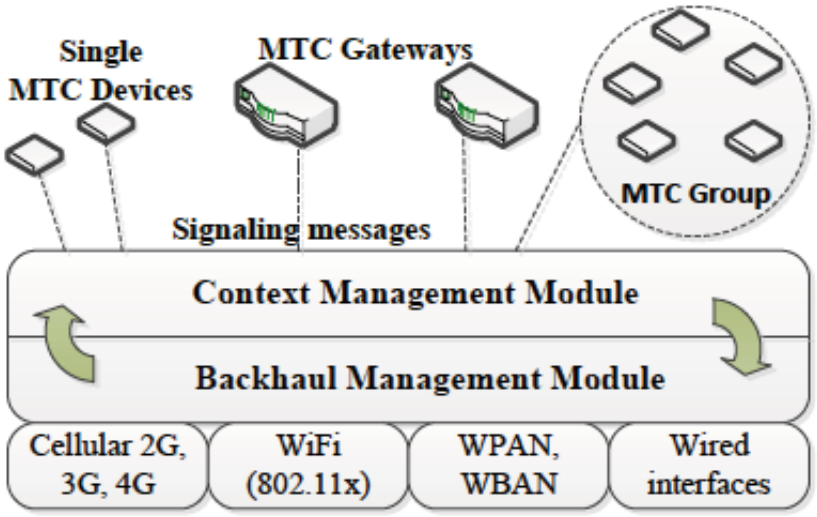

Figure 4.20: Proposed CABM scheme

4.5.2 General design requirements

CABM is aasumed to employ an integrated QoS provisioning approach by handling both H2H and M2M mobile data traffic. By the term H2H, we mean all the well-known service classes being standardized by corresponding specification groups. For simplicity reasons, it is assumed that all H2H services can be mapped to the four QoS classes defined in [181], namely conversational, streaming, interactive and background classes. Furthermore, according to [222], we assume that some MTC services have similar service requirements with the ones defined for the four basic QoS classes and thus can be easily mapped onto them. However, there are many MTC services, which cannot be mapped to known QoS classes. Indicative examples of these novel MTC services are provided at table 4.7.

MTC services' characteristics are determined according to all possible combinations of MTC features. MTC features are defined to provide: a) a structure for the different system optimization possibilities that can be invoked by CABM scheme [222], and b) a MTC-oriented QoS classes categorization [231]. Without lack of generality, we have chosen the following MTC features from the list provided in [222], namely: a) priority alarm, b) time controlled, c) time tolerant, d) high mobility, e) large data transmission, and f) group-based feature. It should also be noted that there is a direct mapping between the six novel MTC services described in tables 4.7 and 4.8.

Table 4.7: Indicative novel (beyond known QoS classes) MTC services

Features / Services	Priority Alarm	Time Controlled	Time Tolerant	High Mobility	Large Data Transmission	Group based
Service 1	✓	✗	✗	✗	✓/✗	✗
Service 2	✓	✗	✗	✓/✗	✓/✗	✓
Service 3	✗	✓	✓	✗	✓	✗
Service 4	✗	✓/✗	✓	✓	✓	✓
Service 5	✗	✗	✓	✗	✓	✓
Service 6	✗	✗	✗	✓/✗	✗	✓/✗

Table 4.8: Examples of novel MTC services & actions by proposed CABM scheme

CABM / MTC Services	BM module actions	CM module actions
Simple Alarm	Accept the MTC service with the highest priority	Store the whole event and dynamic partitioning actions to make smarter future resource allocations
Group-Based Alarm	Accept, assign highest priority and based on CM feedback, reserve even more bandwidth for upcoming similar MTC requests from the same group	Based on group's ID and the acco-mpanied MTC features, predict the bandwidth reservation for the whole MTC service and inform BM
Time Controlled & Tolerant Large Data	Reject the MTC service in overload states	Signaling message to acknowledge context to MTCD and propose a future transmission interval, when system utilization will be lower
Very Time Tolerant	Reject the MTC service in relatively high system utilization but non-overload states	Signaling message to acknowledge context to MTCD group and broadcasting message to all neighboring MTCGs
Group-Based Background	If accepted and based on CM feedback, reserve proactively even more bandwidth for upcoming similar MTC requests from the same group	Based on group's ID and the acco-mpanied MTC features, predict the bandwidth reservation for the whole MTC service and inform BM
Real-Time Small Data	Accept MTC service even in higher system utilization states	In cases of group-based transmission and/or high mobility, consult context history to decide whether to be accepted and re-determine thresholds

For example, services 1 and 2 of table 4.7 (see also simple alarm and group-based alarm services in table 4.8) are very high-priority services (even higher than conversational class), so CABM should treat them as a special QoS class. In case of group-based alarm MTC service, we assume that a group of MTCDs wants to transmit during a small time interval to the MTCG. A representative real market example case would be a building's safety system equipped with surveillance cameras. When an intruder is detected, the cameras of the building will consecutively transmit streaming video scenes once the intruder passes through the various rooms. From the MTCG's perspective, group-based alarm MTC service requests will arrive one after the other, so CABM has to proactively treat them as a whole service (i.e. intruder monitoring) and not as individual service requests. So, apart from assigning the highest priority, CM module can predict the total bandwidth reservation (based on group's ID) for the whole MTC service and inform BM, which will proactively reserve even more bandwidth for upcoming similar MTC requests from the same group. Similar actions from BM and CM modules are expected for group-based background MTC services (i.e. service 5 of table 4.7).

In case a MTC service has time tolerant and time-controlled features enabled (e.g. service 3 & 4), it should be served in a different manner than H2H background services. For example, a solar or battery powered monitoring device may gather time tolerant data, which however have to be transmitted in a short time period in order to minimize power consumption. Therefore, such services should not be served as background (i.e. long service periods at minimum transmission rates), but rather should be rejected and scheduled to be served at a later time at high data rates. Thus, CABM also achieves the distribution of the traffic load within a given time period and the system avoids congestion/overload states.

After BM's decision for rejection, CM sends signaling messages to acknowledge the context to interested entities such as MTCDs, MTCD groups and/or neighboring MTCGs (in case the high mobility feature is enabled), proposing a future transmission interval, when system utilization will be relatively low. Finally, for real-time small data transmissions (up to some KB), CABM can accept them even in higher system utilization states according to the overall contextual information (i.e. accompanied MTC features).

4.5.3 MTC request admission control (MTCR-AC) algorithm

In order to develop a unified service admission control policy, we have to group the various H2H and M2M services into a number of Common Service Classes (CSC). However, as discussed previously, a number of MTC services cannot be properly mapped to any of the well-known QoS classes. So, the typical H2H services classification scheme has to be extended to include MTC-oriented QoS classes. An indicative classification scheme, including two additional CSCs (i.e. CSC 1 and CSC 6) is shown in table 4.9. The first CSC

includes all the alarm MTC services, which require to be served with the highest priority, while the last CSC includes all the time tolerant and time-controlled MTC services, which require to be served in a different manner than H2H background services.

Table 4.9: Indicative service classification

Common Service Classes	Predominant Service Feature	Capacity Partitions
CSC 1	Priority Alarm	P1
CSC 2	Conversational	P2
CSC 3	Streaming	P3
CSC 4	Interactive	P4
CSC 5	Background	P5
CSC 6	Time Tolerant // Time-Controlled	P6
All CSCs		Common Partition

In order to provide a predetermined QoS (in terms of acceptance rate) to the various H2H and M2M services, the already described capacity partitioning concept is adopted (see previous sections). Therefore, the available backhaul capacity is divided into partitions, each one corresponding to a specific CSC. Depending on the maximum offered traffic load intensity for each partition and for a given blocking probability, the sizes of the partitions are calculated based on the Erlang B formula or, if required by the nature of services, by following a more precise approximation such as the one described in [232]. After determining the size of the partitions, any residual capacity forms an additional common pool partition, which can be equally utilized by all the CSCs as shown in table 4.9.

Regarding the proposed MTCR-AC algorithm, the process is as follows: upon the arrival of a service request, the CM module determines its service class by analyzing the relative context. The BM module, based on the available context information, may temporarily postpone a CSC from being served, thus leading to a "context based rejection", while the CM module is informed accordingly. If this is not the case, MTCR-AC sequentially checks if the required resources can be allocated to the incoming service request from the respective partition or, if this is not possible, from the commonly shared partition. If there is adequate capacity available, then the service request is accepted. Else, if the resource allocation process is not completed successfully, a "capacity based rejection" occurs. The operation of the MTCR-AC algorithm is illustrated in figure 4.21.

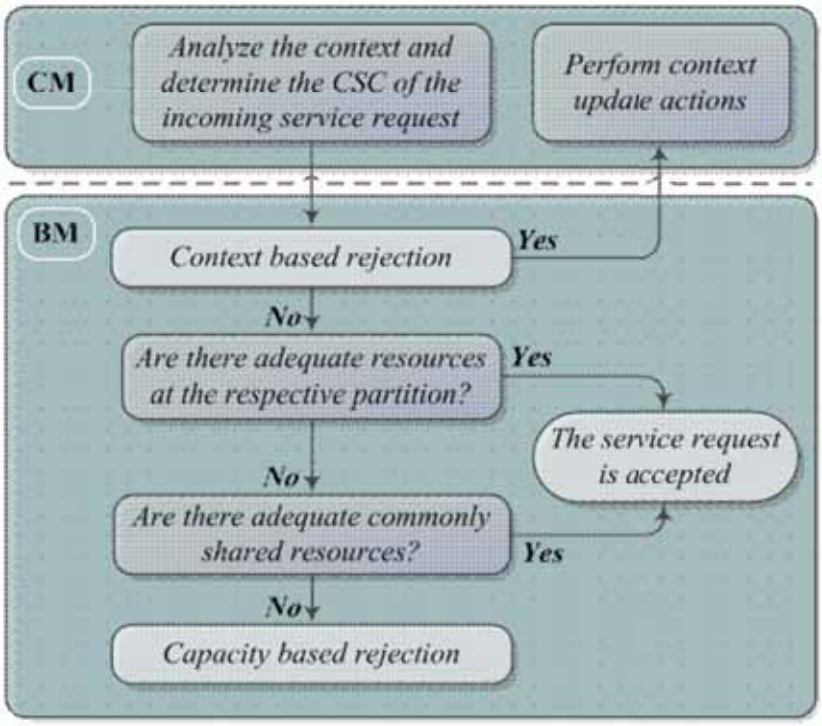

Figure 4.21: Flowchart of MTCR-AC algorithm

4.5.4 Performance evaluation results

In this section, the performance of the proposed CABM scheme is evaluated through event-driven simulation written in C/C++. Service requests are assumed to arrive according to a poisson process, while their duration is exponentially distributed. Each service request is related to a CSC and therefore mapped to a specific partition. The backhaul capacity is partitioned according to the composition of the offered traffic load and by taking into account the available context information. The evaluation of CABM is performed in comparison to a typical UMTS Service Admission Control (USAC) scheme, which is based on a four classes categorization system [181], while on the other hand CABM is based on the extended classification of table 4.9.

Regarding the simulation scenario, a mixed H2H and M2M incoming traffic load is assumed, which is composed by services of the 1st CSC (10%), 2nd CSC (40%), 5th CSC (40%) and the 6th CSC (10%). Services of the 1st CSC require to be served with an extremely low blocking probability, upper bounded to 0.1%, while conversational services of the 2nd CSC require a blocking probability lower than 1%. Background traffic of the 5th CSC is served in a best effort manner, while time tolerant/time controlled traffic of the 6th CSC can be served when the system is not congested.

In USAC case, due to its limited QoS classification system, the 1st and the 2nd CSCs are merged to the highest priority class of 3G and beyond systems, which is the conversational class (i.e. 2nd CSC in the extended classification system). Similarly, the 5th and the 6th CSC

are merged to the lowest priority class of 3G and beyond systems, which is the background class (i.e. 5th CSC). As a result, USAC is not able to provide the required QoS differentiation and this is clearly depicted in figures 4.22 and 4.23. Conversely, as shown in the same figures, CABM is able to provide the required low blocking probability to the Priority Alarm services and differentiate them from the conversational services. Furthermore, by rejecting the services of the 6th CSC, CABM allows them to be served properly at a later time interval, while simultaneously reduces significantly the blocking probability of the H2H background services (5th CSC).

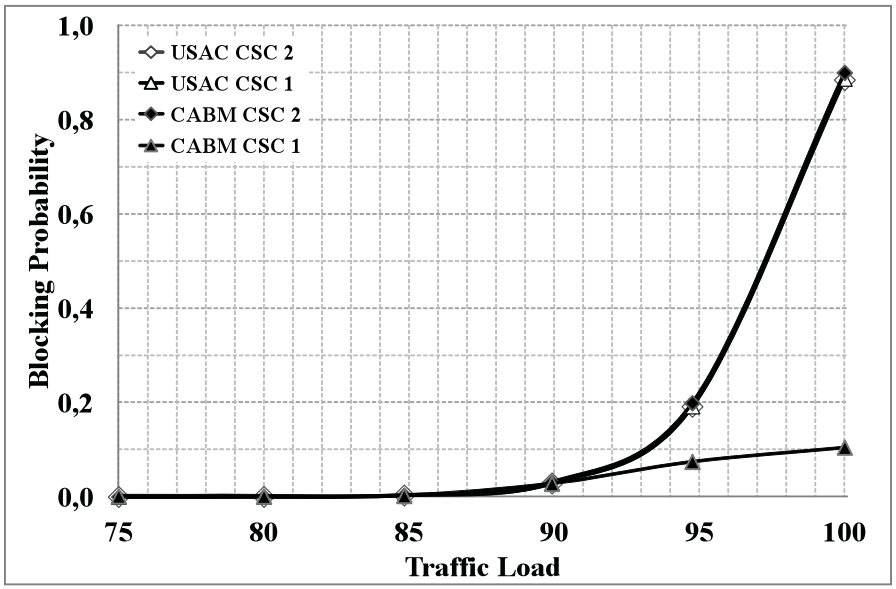

Figure 4.22: Blocking probabilities for services of the 1st and 2nd CSCs

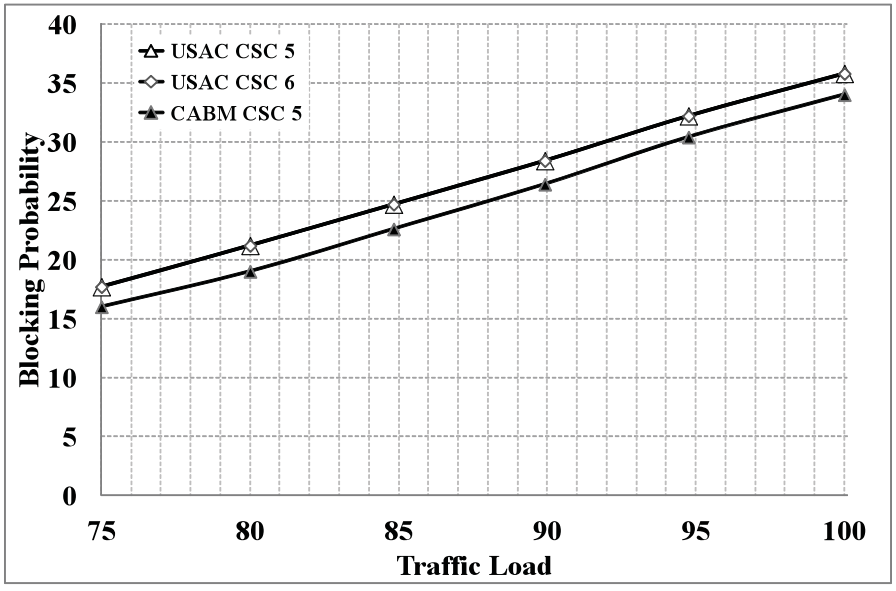

Figure 4.23: Blocking probabilities for services of the 5th and 6th CSCs

4.6 Summary

In this chapter, the Integrated Services Router (ISR) concept was introduced and studied, which is expected to boost mobile and fixed networking systems' convergence research in the upcoming years. Various real market network deployment scenarios were studied and the research problem of context aware resource management for various types of ISRs (i.e. small cell gateways, MTC gateways, etc) was investigated. Furthermore, three novel context aware admission control schemes were proposed, which are applicable in ISRs and employ integrated services QoS provisioning, while one of those schemes (i.e. ISAC), is part of CA-FEI framework as this was introduced in chapter 3. Performance evaluation results have shown that being aware of both 4G HetNet environment's radio and backhaul resources, a context aware resource management module can enhance overall system's key performance indicators such as overall system capacity/utilization and all services' QoS provisioning. Conclusively, it is shown that a converged 4G HetNet deployment setting consisting of both mobile and fixed networking subsystems can better deal with emerging and up-to-date resource management research problems.

CHAPTER 5

CONTEXT AWARE RESOURCE MANAGEMENT IN MOBILE/HYBRID CLOUD INFRASTRUCTURES

According to figure 1.1 and regarding the three main architectural pillars of the current thesis, the scope of the research being undertaken in this chapter refers to the respective outlined area at the right hand side of the figure (i.e. mobile/hybrid cloud infrastructures). For these types of infrastructures, novel context aware resource management schemes are proposed, while the role that cloud computing (CC) technology can play in the evolution of context aware mobile and wireless networking research (CAMoWiN) is the main field of the study. More specifically, at the introductory section, a typical mobile cloud computing (MCC) and a hybrid cloud computing (HCC) environment are described. In section 5.2, the MCC/HCC resources management problem is formulated and the assumed system models are provided. Moreover, an overview of the related works in the international literature outlines the innovative parts of the proposed context aware resource management schemes. In section 5.3, a mobile cloud resources provisioning (MCRP) scheme is proposed, which is applicable for a typical MCC environment. Section 5.4 introduces an IaaS request admission control (IRAC) scheme, which provides a customizable infrastructure sharing approach for HCC environments. The performance of both MCRP and IRAC schemes is evaluated through simulation results, while section 5.5 summarizes and concludes the chapter.

5.1 Introduction

As an emerging information technology, Cloud Computing (CC) has plethora of advantages allowing IT organizations to focus on their business innovation rather than on building larger data centers to achieve scalability goals [233]. In a nutshell, CC paradigm can: a) offer a seemingly unlimited pool of computing resources, b) reduce operational expenditures for organizations, c) reduce general costs by following a pay-as-you-go business model for providing software, platform, infrastructure etc. as a service (*aaS), d) increase IT flexibility and improve IT business processes efficiency boosting for research breakthroughs and innovative business models invention, e) reduce time-to-market processes for new products and services by simplifying the various time-consuming hardware provisioning, purchasing and deployment procedures f) have great impact on non-ICT industries and economies, g) be available to a large number of end users with very disparate needs, h) provide means of sustainable economic growth by offering equal opportunities to start-up companies and SMEs to claim their own market share in a very competitive environment, and i) provide means of

energy efficiency by reducing total carbon emissions caused by ICT, j) boost mobile and wireless networking industry by introducing mobile cloud computing (MCC) concepts.

Based on the well-known service models defined by NIST [234], the most important resource sharing scenarios can be categorized at levels of infrastructure sharing (IaaS), software sharing (PaaS) and application sharing (SaaS). In addition to these, business process sharing scenario can be applicable to many diverse industries and emerging economies namely e-government, e-health, e-logistics, e-agriculture, e-education, e-media, e-banking, e-manufacturing, e-tourism etc. These e-* business scenarios can take advantage of the given potential to aggregate and analyze data from multiple sources creating thus economies of scale. Virtualization is a technology that forms the foundation of CC, as it abstracts away the details of physical resources and provides corresponding virtualized ones for high-level applications on-demand [235]. In figure 5.1, the encapsulation rationale regarding the general CC stack defined by NIST and the main virtualization components, is presented. More specifically, applications are encapsulated within an operating system (OS), which itself is encapsulated within a virtual machine (VM). The VMs are then mapped to physical resources layer, i.e. each VM is assigned to a hypervisor, which is located on a physical server. The hypervisor layer is used to present the abstraction of the underlying physical resources (such as CPU, storage, network) to multiple VMs and thus allows the former to be subdivided and shared. Furthermore, hypervisors maintain adequate partitioning to ensure that a VM cannot access the resources of another VM [236]. As a result, for each type of physical resource, an integrated infrastructure pool is created, which can include physical resources residing in different geographical areas, even though legislative restrictions render obvious the need for dynamic security solutions in public cloud deployments [237].

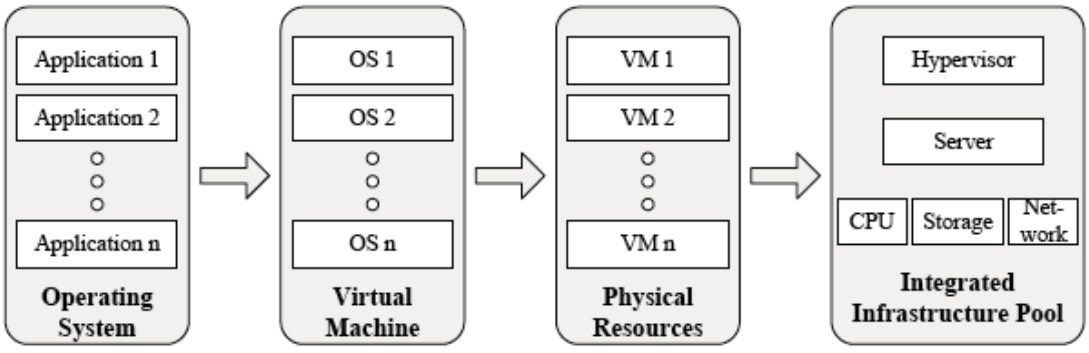

Figure 5.1: General cloud computing stack and virtualization components

The integrated infrastructure pool is a continuously changing quantity and thus capacity (i.e. resources) management becomes crucial in such kind of virtualized environments. Towards this direction, maximizing the utilization of computing resources is a major prerequisite and

thus server consolidation techniques and dynamic (re)-allocation of VMs based on workload changes are required [238]. Intelligent allocation of physical resources for managing competing and diversified (in terms of priority) users' demands can be realized via policy-based management techniques, too [239]. Security is another issue of major importance, which can be confronted by means of VM isolation techniques. The security requirements for cloud providers necessitate corresponding isolation at physical infrastructure level, too. This means that not all parts of the integrated infrastructure pool can accept all IaaS requests, incurring thus even more restrictions in the resource management problem. In these cases, policy-based security management techniques can provide proper decisions about outsourcing information technology services in hybrid CC environments [234] [239]. Finally, making optimal decisions about the extent of outsourced resources (i.e. percentage of non-private resources in the integrated infrastructure pool) can incur additional trade-off dilemmas about ways to deal with abrupt and short-term excessive workloads [240]. Summarizing the basic capacity management objectives that a virtualized cloud environment should satisfy, these are [235]: a) load balancing, b) adaptive resource allocation, c) flexibility, d) security, e) fine-grained resource control, f) continuous resources availability, g) high system utilization performance, and h) deal with unreliably and unpredictably excessive computing resources.

During the last years, CC technology is boosting research innovations for mobile/wireless networking, while a fusion of these heterogeneous research fields is nowadays revolutionizing the ways of communication and computation research trends. As a result, the newly founded Mobile Cloud Computing (MCC) research area is inspired by the notion of complete networking and computing environments' integration and consequently efficient MCC resource management frameworks should be designed and developed that simultaneously take into consideration both: a) wireless/radio access resources pool aiming at always-best connectivity contexts and b) computing resources pool for data processing/storage aiming at flexible virtualized infrastructure sharing solutions. In this chapter, a mobile cloud resources provisioning (MCRP) scheme for MCC environments and a IaaS request admission control (IRAC) scheme for hybrid cloud computing (HCC) environments are proposed.

5.1.1 Mobile cloud computing (MCC) environment

As long as the integration of computing and networking environments is continuously boosted with the assistance of innovative ICT architectures and technologies, implications of cloud computing (CC) paradigm, which are applicable in mobile and wireless networking area are increasingly gaining ground [23]. Indeed, mobile cloud computing (MCC) is an emerging research area and is introduced as the integration of CC into the mobile environment [159]. The ultimate vision is to fulfil the dream of providing "information at everyone's fingertips anywhere at anytime" and as computation capabilities of mobile

terminals (MTs) will always be a compromise, MCC aims at efficiently using CC techniques for data storage and processing on MTs, thereby reducing their limitations [127] [160]. Other major MTs-related technical restrictions are short battery lifetime, varying wireless channel conditions and high network latency, all of them hindering the remote display functionality of cloud applications on mobile devices [161].

Going back to the late '90s, when 3G networks were standardized, a hierarchical cell structure model was proposed in order to better cope with the next generation mobile networking continuum challenges. More specifically, pico, micro, macro and global cell concepts were introduced based on cells' geographical expansion and till now, this hierarchical cell structure is considered as the basis for architectural mobile networking innovations [241]. Nowadays, inspired by this widely accepted abstract paradigm, the future computing continuum is expected to embrace distant cloud infrastructures, proximate cloudlet infrastructures, communicating objects and smart devices.

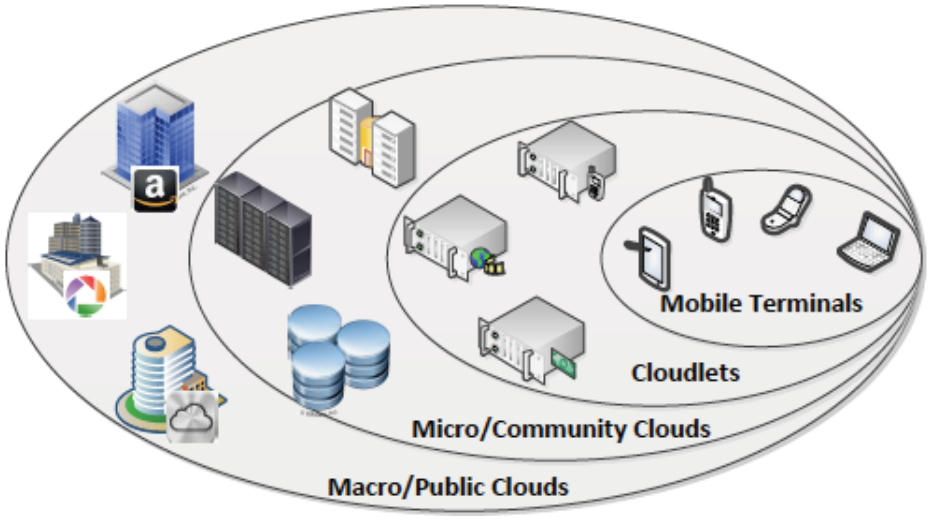

Figure 5.2: MCC hierarchical structure

Table 5.1: MCC hierarchical structure attributes

CC Hierarchical Structure	Expansion	Applicability	Pros & Cons
MacroCloud	Public Clouds	Amazon/Apple/Google/ Microsoft large datacenters/datafarms	(+) high availability, (-) high WAN latency
MicroCloud	Community Clouds	e-gov/health/learning/ banking/commerce/ logistics datacenters	(+) targeted industrial/institutional sectors applicability (-) high WAN latency
Cloudlet	Local/Private Clouds	cafes, shopping malls, airports, stadiums, museums, campuses, train stations, etc.	(+) low latency, real-time services (-) context-aware applications range restrictions
Mobile Terminal	Personal MT Computing	smartphones, tablets, laptops, etc.	(+) security, privacy (-) battery power restrictions

As shown in figure 5.2 and table 5.1, the MCC hierarchical structure can be composed of four different computation layers. Today's MTs (i.e. laptops, smartphones, tablets, PDAs) can provide advanced computing capabilities (compared to the past decade), but the inherent problems of battery power restrictions are inevitable. Moreover, with the explosion of mobile applications market, the average mobile user demands for computing/storage power is much higher than the one that can be supported by an average MT and this gap is continuously growing [127]. As a result, mobile users need to access servers located in virtualized CC infrastructures in order to meet their increasing functionality demands. A cloudlet infrastructure consists of a cluster of servers well-connected to the Internet and available for use by nearby MTs. A cloudlet can be contained in a MCC hotspot together with a wireless access point comprising thus a datacenter-in-a-box concept [158]. These MCC hotspots can be placed in cafes, shopping malls, airports, stadiums, museums, city squares, campuses, train stations, etc. In these environments, mobile users may meet the demand for real-time interactive responses for specific-purpose context-aware mobile applications by low-latency, one-hop and high-bandwidth wireless access to the cloudlet. The problem with cloudlets is their restricted computing/storage capabilities, taking into account that they have to (by priority) meet the QoS demands of numerous users for specific-purpose context-aware mobile applications in a restricted geographical space. Consequently, for general e-* related MCC services (i.e. e-government, e-health, e-banking, e-commerce, e-learning, etc), the access to a community cloud infrastructure would be more appropriate for mobile users requesting corresponding specific e-* mobile applications. A community cloud or micro-cloud (cf. microcells in mobile networking continuum) is controlled and used by a group of industrial/institutional organizations, which have common or shared financial, security and legal objectives [34]. In spite of their high availability and targeted applicability to specific mobile users' demands, high WAN latency is an inevitable trade-off. The same problem is valid for public clouds or macro-clouds (cf. macrocells in mobile networking continuum). The main difference is that Amazon/Apple/Google/Microsoft etc large datacenters can provide virtually infinite computing/storage capabilities for any type of MCC service.

5.1.2 Hybrid cloud computing (HCC) environment

According to NIST specifications [234] and other recent CC-related survey papers such as [235], the basic cloud deployment models are four, namely a) private, b) community, c) public and d) hybrid cloud. A private cloud infrastructure is operated solely for an organization. In other words, each organization, enterprise, institute etc. that adopts the CC technology, builds its own data center investing its own capital for avoiding restrictions of network bandwidth, security and privacy exposures, legal requirements etc. that using public cloud services might entail [237]. Instead of the exclusive utilization of infrastructure and

computational resources by a single organization, community cloud is controlled and used by a group of organizations, which have common or shared financial, security and legal objectives. For example, several state ministries can form an e-government community cloud, several hospitals and clinics can form an e-health care community cloud, many wholesale traders and retail companies can collaborate in an e-logistics use case and many schools, academic and research institutions may take advantage of e-learning opportunities. As a result, resource sharing among communities can enhance computing/storage capacity and thus provide new resource management related research challenges [242]. The same is also valid for public cloud infrastructures, which can provide elastic and cost effective means to deploy IT solutions especially for organizations that don't have adequate funds to acquire and maintain their own large data centers. However, they have some major drawbacks such as inadequate security and privacy support, coarse-grained access control mechanisms and unreliable QoS provisioning. Conclusively, hybrid cloud infrastructures (i.e. composition of the previous three) appear as the best solution, as they can bring out the various strengths by simultaneously limiting the corresponding weaknesses of the pre-referred cloud deployment models. In fact, cloud end users can typically outsource non-business critical information by processing it in a public cloud, while keeping business-critical services and data under their control. Seen from a resource management perspective, whenever a private cloud cannot handle excessive workloads, an appropriate portion can be migrated for processing in community or public clouds. Furthermore, community-related IaaS requests have to be processed in a community cloud, as data available only at community cloud data centers may have to be combined and analyzed.

Multiple parameters have to be taken into consideration for optimum resource (i.e. capacity) management. For this chapter's problem formulation purposes, three basic real-market applicability scenarios have been chosen, namely: a) In-House oriented Cloud Infrastructures (IHCI), b) Community-Sharing Cloud Infrastructures (CSCI), and c) Outsourcing oriented Cloud Infrastructures (OCI). Given the fact that a cloud provider has a clear understanding of the potential of each federation decision, the most profitable and convenient scenario can be chosen according to the context. At the remainder of this section, the pre-mentioned scenarios are described according to some qualitative performance indicators as indicated in table 5.2.

Table 5.2: Qualitative performance indicators and general hybrid cloud infrastructure scenarios categorization

	In-House oriented Cloud Infrastructure (IHCI) scenario	Community-Sharing Cloud Infrastructure (CSCI) scenario	Outsourcing oriented Cloud Infrastructure (OCI) scenario
Representative capacity utilization percentages (private-community-public)	80 – 15 – 5	30 – 60 – 10	10 – 30 – 60
Type of IaaS requests	business-critical	community-critical	non-business critical
Security support provisioning	high	depends on the type of the community group	relatively low
Level of trust among cloud end users and IaaS provider	high	moderate	relatively low
High-priority groups QoS satisfaction level	optimal	moderate to high	relatively low
Costs for outsourcing	low	medium	high
Operational expenditures	high	medium	low
e-business use case applicability	e-government, e-health, etc.	e-logistics, e-media, e-education, etc.	Start-up companies, SMEs
Cloud provider's energy gains	good only in under-utilized conditions	depends on the size of community partition	relatively high

5.1.2.1 In-House oriented Cloud Infrastructures (IHCI)

In this scenario, it is assumed that organizations maintain relatively large proprietary infrastructures for their own use mainly because they possess/process huge amounts of sensitive data, which is unlikely to migrate to community/public clouds due to privacy and security reasons. Private CC resources partition comprises the largest portion of the integrated pool and thus only non-business-related IaaS requests are migrated to the public cloud with the prerequisite that this is done only if costs for outsourcing are lower than processing the requests in-house. For over-utilization conditions, community clouds can offer a satisfactory alternative so that high-priority users can enjoy high QoS, while for under-utilization cases the business-level profits cannot be optimum, even though energy-efficient capacity management is feasible (e.g. by maximizing the number of physical machines shut down when not in use). In EU co-funded PASSIVE project's context [243], undertaken research deals with e-government use case, which can be matched with IHCI scenario. More specifically, a ministry can maintain an IHCI utilizing it for its security-sensitive information handling operations. Minister's office personnel is a high-priority group and thus corresponding IaaS requests from trusted cloud end users should be served by private partition. Other guests and inferior employees may be served with lower priority, while non-governmental related processing can be migrated to public cloud in order to ensure by all means availability of critical private CC resources. Finally, a governmental community cloud's resources can be exploited in cases

where information from several ministries should be combined (i.e. an employee at the ministry of labour can store/retrieve data in the community cloud so that other employees at ministries of finance, healthcare, interior affairs etc. can have access, too).

5.1.2.2 Community-Sharing Cloud Infrastructures (CSCI)

As already stated, non-ICT industries and economies are continuously growing, taking advantage of CSCI. For example, in an e-logistics use case, import/export companies, wholesale traders and retail enterprises which have common business goals and sharing profits in the value chain, can have access to the same community cloud, whose cloud end users enjoy a satisfactory level of trust with IaaS provider compared with OCI scenarios described in the next subsection. As a result, each enterprise doesn't have to maintain and pay OPEX for a large proprietary infrastructure such as IHCI scenario at the expense of paying a relatively smaller fee for its participation to the community cloud. CSCI scenario can also be applicable in other e-business use cases such as e-agriculture, e-media, e-education, e-banking, etc. According to each use case's context, capacity utilization percentages may differ and thus each organization can define its own business-level objectives giving thus feedback to lower-level resource management modules for optimized solutions regarding the qualitative performance indicators presented in table 5.2.

5.1.2.3 Outsourcing oriented Cloud Infrastructures (OCI)

This scenario assumes small private CC partitions and it perfectly suites with start-up organizations' and companies' objectives. More specifically, start-up SMEs can invest a relatively small capital for their business initialization steps, acquiring a small private cloud. OPEX costs can also be minimized offering simultaneously the opportunity to SME's managers about deciding the time that their private virtual infrastructure will have to grow in order business-level objectives to be met. For example, if costs for outsourcing become considerably higher than corresponding OPEX incurred from maintaining computational resources in-house, then private CC partition has to be increased. In OCI, it is also assumed that need for QoS and security support provisioning is not vital and the majority of IaaS requests are non-business critical. For example, compute-consuming research experiments and simulations can be processed in public clouds, while few IaaS requests by high-priority cloud end users can be processed in private or community clouds. The proposed IRAC scheme presented in section 5.4, apart from its optimized capacity management capabilities, can enforce modifications to business-level policies mainly in terms of changing capacity utilization percentages among the three abstract partitions presented in section 5.2.2 (see figure 5.5).

5.2 Mobile/Hybrid Cloud Computing Resources Management

In this section, the assumed system models are presented and the research problem of context aware resource management in mobile/hybrid cloud infrastructures is formulated, while an overview of the related works found in the international literature is provided, too.

5.2.1 MCC resources provisioning problem

In order to formulate the MCC resources provisioning problem, a clear MCC paradigm realization is needed. In figure 5.3, five MCC reference use cases are presented. In a nutshell, different kind of MTs (i.e. cell phones, smartphones, tablets, PDAs, laptops, etc) have first to be connected to one of the available heterogeneous radio access technologies (e.g. LTE, WiFi, WiMAX, UMTS, HSPA, femtocell, WPAN, etc). Once a MT is connected to a wireless access technology, a plethora of different alternatives arise, regarding the ways that efficient partitioning of the computation workload of a mobile application can be achieved between the MTs and the cloud infrastructure. Dynamic partitioning of mobile applications can offer fine-grained flexibility on the "what to process/store where" problem in order to cope with the MCC heterogeneity challenges (e.g. workload, network, device, etc) [162].

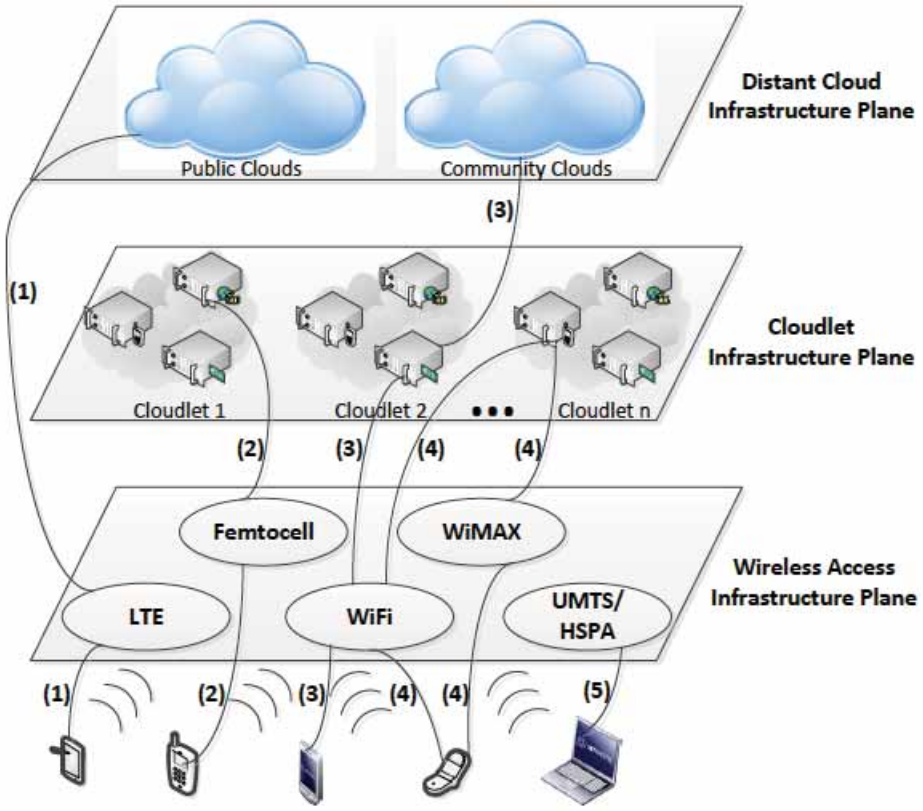

Figure 5.3: MCC reference use cases

More specifically, in use case (1), a MT is connected to a LTE network and uses computing resources from a public cloud infrastructure (e.g. gmail, dropbox, icloud, etc) in order a

mobile cloud application to be executed. In MCC context awareness use case (i.e. use case 2), a MT is connected to a femtocell and uses computing resources from a cloudlet infrastructure. For example, a user enters a shopping mall area and wants to know about all the events, discounts and social networking information related with his/her presence in this geographical area. If he/she is a regular visitor at the shopping mall, a clone of his/her smartphone can exist in one of the servers of the cloudlet infrastructure and thus real-time context-aware mobile applications could run both on the MT and the cloudlet. Furthermore, there is the case in which the dynamic partitioning of a context-aware mobile application can include computation offloading to a distant cloud infrastructure, too (i.e. use case 3). For example, for a e-gov mobile application, some "heavy" non real-time operations can be executed in a distant e-gov community cloud, some other real-time operations in the cloudlet before all data are displayed at the MT's side. In the MCC mobility management use case (i.e. use case 4), a MT is initially connected to a WiFi hotspot and cloudlet computing resources are used. While the mobile application is running, the user moves out of the coverage of the WiFi and thus seamless connectivity with another RAT (e.g. WiMAX) has to be achieved without violating QoS constraints. Finally, the fifth variant considers the trivial MCC use case, which assumes that complete execution of a mobile application on a MT is preferable compared to previous computation offloading scenarios [244].

Table 5.3: Mapping of state-of-the-art MCC challenges to MCC use cases

MCC Challenges / MCC Use Cases	Limited bandwidth/ high latency	Mobility/ Connectivity	Context-aware MCC services/ Business logic	Security/ Privacy/ Trust	Energy efficiency
(1) Distant Cloud Connection	√	L	-	√	L
(2) MCC Context Awareness	L	L	√	P	-
(3) Optimal MCC resources distribution	L	L	L	√	√
(4) MCC mobility management	P	√	L	P	L
(5) Autonomous MT	L	L	-	-	√

Summarizing the MCC reference use cases depicted in figure 5.3, it is obvious that there are many more combinations between the three infrastructure planes. The classification of all possible combinations into five reference use cases was done in order an appropriate mapping with state-of-the-art MCC challenges to be realized (see table 5.3). According to recent

international literature [23] [104] [127] [159] [160] [244], the major MCC challenges are: a) limited bandwidth and high WAN latency, b) wireless access availability and mobility management, c) efficient context-aware MCC services provisioning and business logic issues, d) security, privacy and trust issues and e) energy efficiency issues. In table 5.3, the grade of impact that MCC challenges have on the assumed MCC use cases is qualitatively evaluated. Hence, in "distant cloud connection" use case (i.e. use case 1), the major problem is the high WAN latency incurred not allowing QoS constraints of real-time services to be met. Security, privacy and trust-related issues are also a major concern, because users often do not have an entire view of the ways that their data are processed in distant cloud infrastructures. In MCC context awareness use case (i.e. use case 2), the major concern is on inventing ways that mobile applications can be useful to users exploiting location-based information and spatial augmented reality concepts boosting thus the MCC market towards introducing opportunities for new business players and models. In "optimal MCC resources distribution" use case (i.e. use case 3), the major challenge is to dynamically partition computational operations of mobile applications between MTs, cloudlets and distant cloud infrastructures in order to optimize specific key performance indicators such as energy, security, QoS, etc. The fourth use case deals with mobility management in MCC and as such, seamless connectivity and handover frameworks have to be carefully designed [104]. Finally, in "autonomous MT" use case (i.e. use case 5), energy efficiency issues are of major concern in order to extend the MT's battery power autonomicity.

According to the problem formulation descriptions, there are numerous different alternatives in the MCC resources provisioning problem. In any case, there is a need to allocate both networking and computing resources simultaneously following joint resource management principles [146]. That is, wireless heterogeneous networking resources allocation problem has to be stressed in conjunction with the cloud computing resources allocation problem and not as two independent sub-problems.

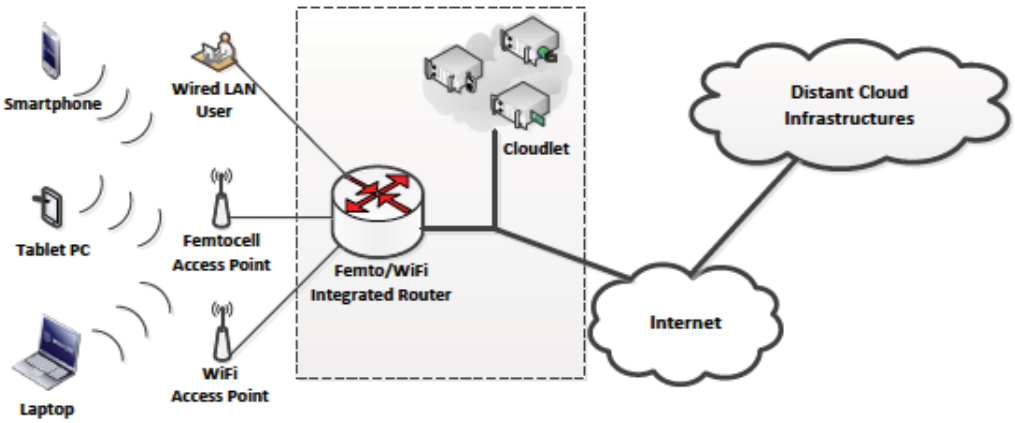

Figure 5.4: System model for a MCC environment

In figure 5.4, the assumed system model for a typical MCC environment is presented. The geographical area, in which this system model can be applied, is restricted and refers to scenarios of shopping malls, city squares, stadiums, museums, airports, train stations, campuses etc. In this kind of places, cloudlet infrastructures are soon going to be deployed in a wide range. These cloudlet infrastructures can be deployed much like WiFi and femtocell access points today and it would be relatively straightforward to integrate cloudlet and WiFi/femtocell APs hardware into a single easily deployable entity [127]. In [158], this integrated infrastructure (i.e. wireless access and cloudlet infrastructure) is introduced as "MCC hotspot" notion. In the assumed system model of figure 5.4, the two pre-referred infrastructures can also be physically separated (i.e. in different physical spaces) but in any case are virtually and functionally integrated (see figure's 5.4 outlined area).

As shown in figure 5.4, mobile users request for MCC services using their MTs such as smartphones, tablet PCs and laptops. These MTs can be connected to more than one heterogeneous radio access technologies (RATs) such as LTE femtocell and WiFi access points. Traffic can also be generated by wired LAN users. The resource management module, which resides at the femto/wifi integrated router has to dynamically partition the available backhaul traffic according to the principles proposed in chapter 4 of the current thesis (i.e. DSAC and ISAC schemes). The resource management module, which resides at the cloudlet infrastructure, has to determine an integrated hybrid CC resources pool at any time instance and efficiently allocate computing resources using virtual partitioning methods and user-oriented, customizable infrastructure sharing approaches such as the ones described section 5.4 for proposed IRAC scheme.

5.2.2 HCC resources provisioning problem

As shown in figure 5.5, the main resource/capacity management problem is to dynamically determine the integrated hybrid cloud infrastructure resources pool at any time instance. Given the fact that all three abstract partitions need to coexist in a typical HCC environment, an effective design algorithm of a general hybrid cloud infrastructure needs to determine the optimal split between private, community and public resources partitions [235]. The summation of these three partitions at any time instance formulates the whole capacity pool that the IRAC scheme (proposed in section 5.4) will efficiently manage. In figure 5.6, the resource allocation procedure is illustrated via a VM provisioning example. As explained in figure 5.1, virtualized resources can be further mapped to physical resources layer instantiating the integrated hybrid cloud infrastructure resources pool shown in figure 5.5.

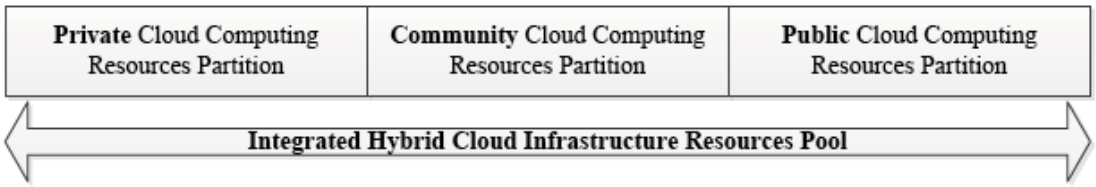

Figure 5.5: The integrated hybrid cloud infrastructure resources pool

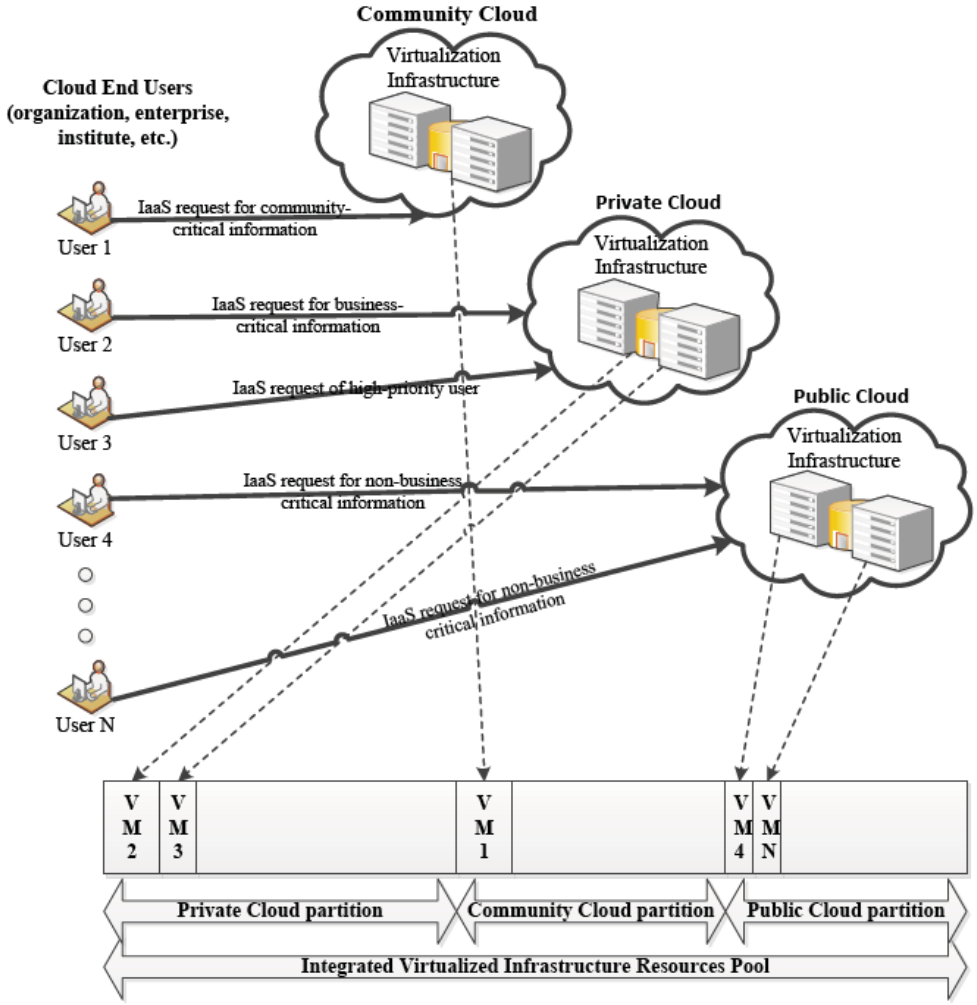

Figure 5.6: System model for a HCC environment

5.2.3 Related work overview

Because of the fact that cloud computing and virtualization technologies are young regarding their research maturity, the majority of the related works found in the international literature are recent. The scope of this chapter's related work includes resource management issues applied in MCC/HCC environments taking into account: a) service differentiation among multiple and diversified user priority groups, b) security restrictions in assigning physical resources, c) overall energy consumption, d) business-level objectives and cloud economics, and e) related federated cloud infrastructures deployment models for e-* use cases.

Regarding (a), plethora of capacity management approaches for virtualized environments have been proposed such as [236] [238] [240] [245] [246] [247] [248] trying to efficiently allocate capacity towards mainly satisfying service differentiation goals. However, their main drawback is that rarely changing infrastructures are assumed and thus cannot effectively deal with over-utilized and unpredictably changing workload situations taking place in HCC environments. Regarding (b), in recent survey-oriented papers like [235] [237] and [249], security and privacy concerns are raised outlining the fact that different security policies in the resource allocation procedures have to be associated with various hybrid cloud service deployments (i.e. in-house virtual/physical resources vs. outsourced ones). In fact, there are some institutional and business sectors, which run more security-sensitive operations like e-government and e-health use cases [250] [251], where end users belonging to the same priority group in terms of QoS, may have different security requirements and consequently different security-related restrictions in parts of the integrated infrastructure pool that have to be taken into consideration, too. In [239], some basic implications of policy-based security management principles are provided, in [248], security is considered as a major factor for the proposed resource management scheme while in [237] general legal requirements for information security and corresponding compliance needs in HCC environments are discussed. Other works considering energy consumption as the major factor are also available [245] [252] [253]. Business-level objectives and cloud economics features in business sharing CC scenarios are investigated in [254] [255] [256]. In these works, ways that high-level business objectives can be mapped to low-level policies implemented in resource management modules are proposed. Finally, cloud federation is a new research trend and thus several research challenges are emerging dealing with ways that resource management modules may be upgraded in order to cope with new multi-parameter optimization requirements [242] [249] [255] [256].

In [253], an extensive presentation of the main autonomic resource management procedures of a virtualized infrastructure is given, namely monitoring, workload prediction, admission control and resource allocation. Monitoring components measure the performance metrics of each incoming IaaS request, categorize them in multiple request classes and estimate requests service demands, while predictors forecast future system capacity situations based on historical data. For the novel context aware resource management schemes proposed in this chapter, the information from the monitoring and predicting components is considered to be known at any time instance and is used as input for the admission controller and the resource allocator. Finally, the abstract resource allocation strategy described in [240] is adopted, which asserts that a priori and a posteriori steps have to be followed.

5.3 Proposed Mobile Cloud Resources Provisioning (MCRP) Scheme

As discussed in the previous sections, the main technical contribution of this chapter lies in the assertion that in the integrated networking and computing continuum, state-of-the-art resource management frameworks and techniques have to be enhanced in order to confront the related research challenges from both networking and computing perspectives simultaneously. Generally, efficient MCC resource management frameworks should simultaneously take into consideration both: a) wireless/radio access resources pool aiming at always-best connectivity contexts and b) computing resources pool for data processing/storage aiming at flexible virtualized infrastructure sharing solutions. In this chapter, a novel mobile cloud resources provisioning (MCRP) scheme is proposed. The aim of MCRP scheme is to jointly handle both the radio and computing resources of the integrated system model as it is depicted in figure 5.4. MCRP is flexible enough to adapt to the various general MCC reference use cases having been described in section 5.2.1. The main novelty feature of the employed MCC Service Admission Control algorithm lies in the fact that it jointly handles radio and computing resources rather than confronting the problem as two independent resource management sub-problems. The performance of MCRP scheme is evaluated via simulation results showing that the assumed context-aware service admission control policies can lead to efficient mobile cloud resources provisioning outcomes.

5.3.1 MCC service classes

Assuming four basic resources (i.e. bandwidth, CPU, memory and HDD capacity partitions) to be managed by MCRP, each new service request (SR) can be expressed as a four-dimensional request vector of the form $[A_{(1,x)}, A_{(2,y)}, A_{(3,z)}, A_{(4,k)}]$, x, y, z, k where $A_{(1,x)}$, $A_{(2,y)}$, $A_{(3,z)}$, $A_{(4,k)}$ are the requirements of the service for each of the four resources. Services that have similar requirements from the same resource are mapped to be served by the same partition.

An example of mapping two different MCC services, namely S_1 and S_2, to multiple resource partitions, is shown in figure 5.7. Both services require similar bandwidth and storage capacity and thus they are mapped to the same partitions $P_{(1,1)}$ and $P_{(4,1)}$. However, the two services have different processing (CPU and Memory) requirements and thus S_1 is mapped to $P_{(2,1)}$ and $P_{(3,1)}$, while S_2 is mapped to $P_{(2,2)}$ and $P_{(3,2)}$ respectively.

Depending on the maximum value of the total offered traffic load intensity for each partition and for a given value of blocking probability, the size of each partition is calculated based on the Erlang B formula or, if required by the type of services, by following a more precise approximation [257]. In a nutshell, the request vector of a service can be corresponded to a partition vector of the form $[P_{(1,x)}, P_{(2,y)}, P_{(3,z)}, P_{(4,k)}]$ x, y, z, k . MCC services with similar characteristics have the same partition vector and belong to the same service class (SC).

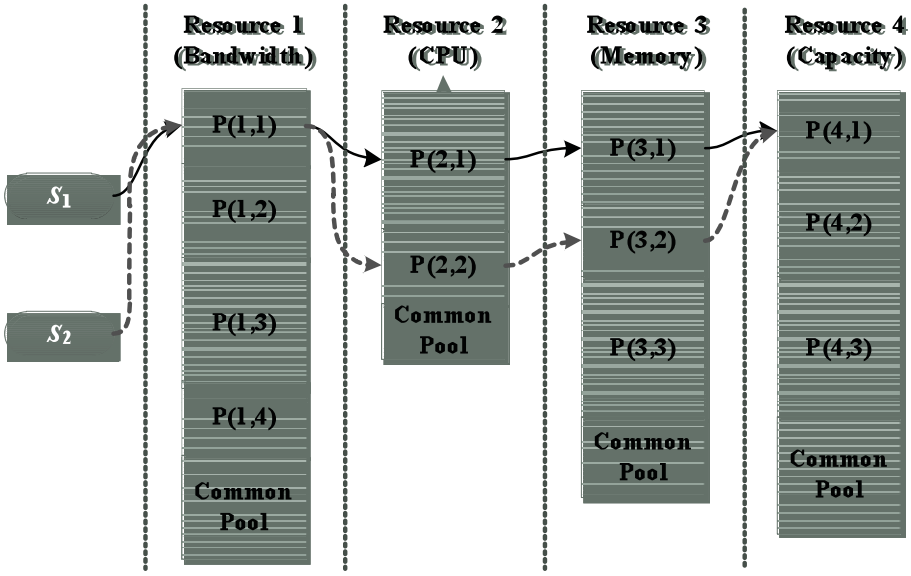

Figure 5.7: Example of mapping MCC services to resource partitions

5.3.2 Partitioning adjustments and limitations

Although the size of the partitions is calculated on the basis of preserving a specific QoS in terms of blocking probability, in the general case one cannot always expect a perfect alignment of the partitions among different kind of resources. This is due to various limitations, which are mainly imposed by the fact that distributed resources have to be managed, which are virtually integrated. For instance, the available bandwidth of the radio interfaces can be constrained by the available bandwidth of the backhaul line that provides access to a distant cloud. Furthermore, the excessive latency of accessing a distant cloud may also exclude delay sensitive service classes by utilizing its computing resources. Consequently, the topology of such a distributed system as well as the technology used in its various components may crucially affect the degree of integration that can be achieved. In other words, the formed partitions may not always have the required capacity, even if the respective integrated resource is adequate to handle the total incoming traffic load.

Based on the above considerations, partitions that belong to the same partition vector [$P_{(1,x)}$, $P_{(2,y)}$, $P_{(3,z)}$, $P_{(4,k)}$] x, y, z, k of the *j*-th service class SC_j may offer different blocking probabilities, if they are independently calculated. In this case, the actual blocking probability B_j of the SC_j service class is lower bounded by the highest blocking probability that each of the individual partitions is able to provide:

$$B_j = \max\left\{B_{(1,x)}, B_{(2,y)}, B_{(3,z)}, B_{(4,k)}\right\},\ x,y,z,k \in N \qquad (5.1)$$

Therefore, for example, there is no point to have low blocking probability at the radio access level, if a MCC service request is going to be blocked due to lack of computing resources. Having this in mind and aiming to preserve a high utilization of system resources, all the

partitions have to be re-examined and subsequently determine which should be reduced in size.

Thus, subsequent to the initial calculation of the partitions, an adjustment phase follows. During this phase, the size of each partition that is allocated to n service classes SC_j, j=1, 2,.., n is downsized so as the blocking probability it offers is not lower than the minimum L of the blocking probabilities of these classes:

$$L = \min\{B_j\}, j = 1,2,..,n \qquad (5.2)$$

The resources that remain unallocated after this phase, if any, are forming a commonly shared partition (common pool) per resource (cf. figure 5.5), which can be accessed by all service classes as long as the limitations described above are met. Otherwise, if a part of these resources can be used only under specific limitations, then this part forms a separate partition. Therefore, the commonly shared partitions may be further divided, if necessary, in order to form smaller homogeneous partitions. These partitions are utilized by MCRP in order to increase system's capacity utilization by absorbing small traffic load distribution variations.

5.3.3 MCC service admission control algorithm

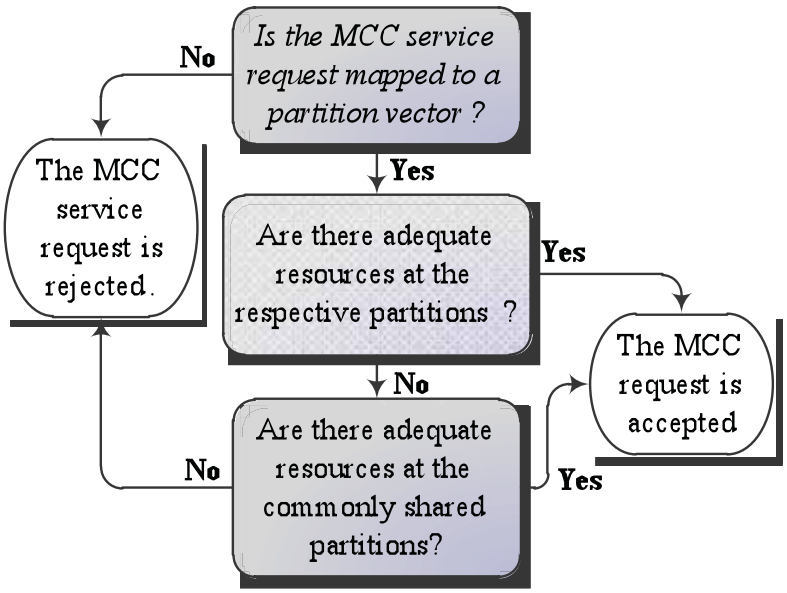

Figure 5.8: Flowchart of the proposed MCRP algorithm

Upon the arrival of an MCC service request that belongs to the j-th service class SC_j, MCRP has to check its partition vector. If the MCC service request is not mapped to a partition vector (i.e. excluded from service), then it is rejected. Otherwise, MRCP sequentially checks if the service request can have the required resources from the partitions included in the partition vector, or if this is not possible, it checks if these resources can be allocated from the

commonly shared partitions. If the resource allocation process is completed successfully, then the service request is accepted or else it is rejected.

5.3.4 Performance evaluation results

In this section, the performance of the proposed MCRP scheme is evaluated through event driven simulation written in C/C++. MCC service requests are assumed to arrive according to a Poisson process, while their duration is exponentially distributed. For comparison purposes, a Complete Sharing Scheme (CSS) is also evaluated as well as a Complete Partitioning Scheme (CPS). CSS is a usual resource sharing approach for Cloud Computing environments [249], where the reserved resources are utilized as commonly shared pools, equally available for every incoming MCC request. On other hand, CPS performs typical complete partitioning of each resource independently, without addressing the need for joint and thus context aware resource management. Finally, the proposed MCRP scheme, being aware of all types of resource partitions at any time instance, can achieve a broader context aware resource management procedure for MCC services.

As shown at figure 5.7, it is assumed that four main resources are shared among the end users, namely bandwidth, processing power, memory and storage capacity. For generality purposes, these types of resources are referred as 1, 2, 3 and 4, assuming that each MCC service requires a number of Basic Units (BU_m, m=1,2,3,4) from each one of them. Consequently, the four-dimensional request vector of each new Service Request i (SR_i) can be expressed as (R_1 BU_1, R_2 BU_2, R_3 BU_3, R_4 BU_4) R_1, R_2, R_3, R_4. For simplicity reasons, three kinds of services are considered for MCRP scheme's performance evaluation purposes: a High Demanding Service (HDS), a Low Demanding Service (LDS) and a Best Effort Service (BES), which require resources that are multiples of a basic request vector of R_V=(1 BU_1, 1 BU_2, 1 BU_3, 1 BU_4). More specifically, the requirement of LDS in resources is R_{LDS}=R_V, HDS has a requirement of R_{HDS} =16·R_V and BES has a request vector of R_{BES} =5·R_V.

The goal is to show that the management of mobile cloud resources should move from CSS schemes to Partitioning Schemes (PS) in order to be able to provide QoS provisioning. However, the employed PS schemes should take into account the peculiarities of a distributed system that is virtually integrated.

In this scenario, it is assumed that the resources of a single cloudlet without any access to a distant cloud have to be shared. The traffic load distribution is set to: HDS 20%, LDS 60% and BES 20%. The blocking probability of both HDS and LDS services should be less than 5%, while the BES service is served as long as there are available resources.

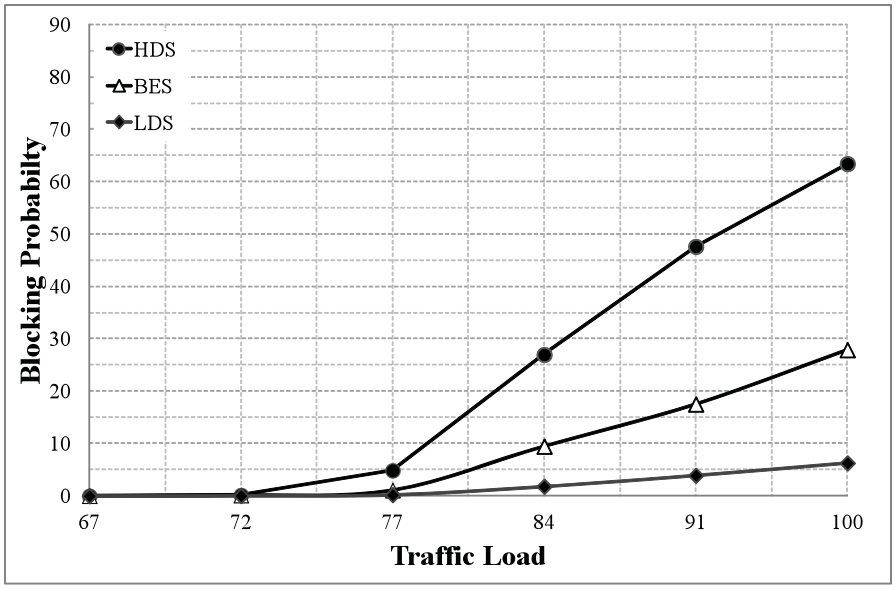

Figure 5.9: Blocking probabilities for the CSS scheme

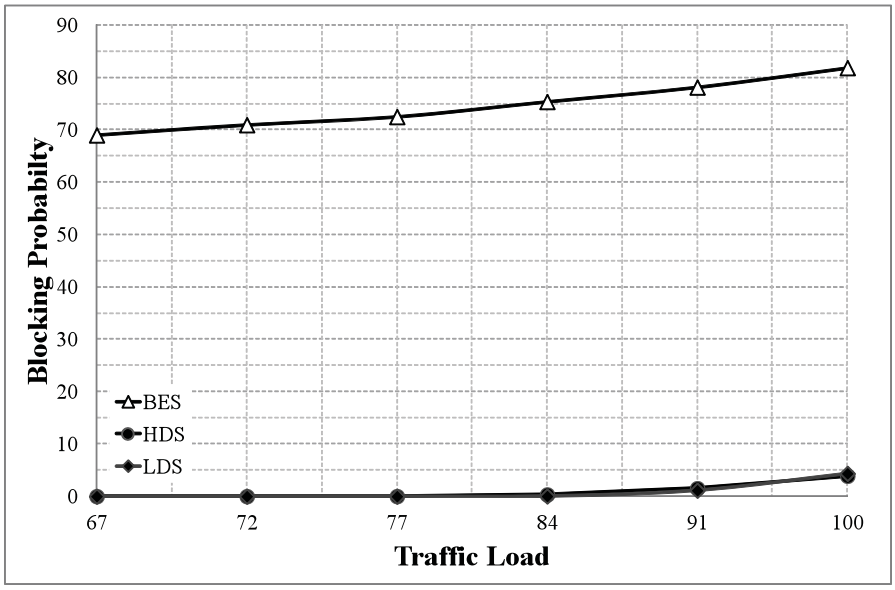

Figure 5.10: Blocking probabilities for the MCRP scheme

As shown in figure 5.9, CSS is not able to provide QoS differentiation and thus the blocking probability of a service is based only on its requirements. The more resources a service requires, the more difficult it is to be accepted. As a result, the blocking probability of HDS services exceeds 5% well before the system is fully loaded. On the other hand, MCRP is able to keep the blocking probability of both HDS and LDS services below the required blocking probability, as shown at figure 5.10, even when the system is fully loaded.

Extending the previous simulation scenario, it is assumed that 15% of the system's storage capacity is derived from a distant cloud. We further assume that the HDS and LDS services are delay sensitive and they cannot tolerate the high latency of a connection to the distant cloud. By following a typical CPS approach, each of the LDS and HDS partitions will include

approximately a 10% of resources, which will never be used due to the limitations at the storage level. On the contrary, MCRP reduces the size of the LDS and HDS partitions, at all resource levels, in order to be adjusted to the available storage capacity of the cloudlet. Subsequently, MCRP adds the released resources to the existing common pool partitions.

As a result, both CPS and MCRP schemes are able to provide the same, increased, blocking probability for the LDS and HDS services, while the extended common pool partitions formed by MCRP provide significantly lower blocking probability for the BES service requests. Thus, MCRP is able to provide better utilization of the available resources and this can be verified at figures 5.11 and 5.12, where the performance of CPS and MCRP schemes is shown respectively.

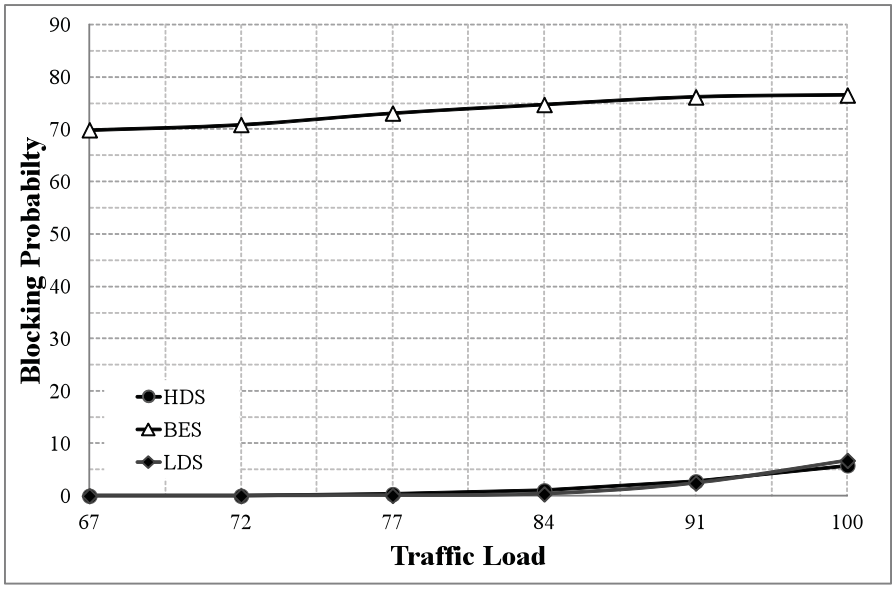

Figure 5.11: Blocking probabilities for the CPS scheme

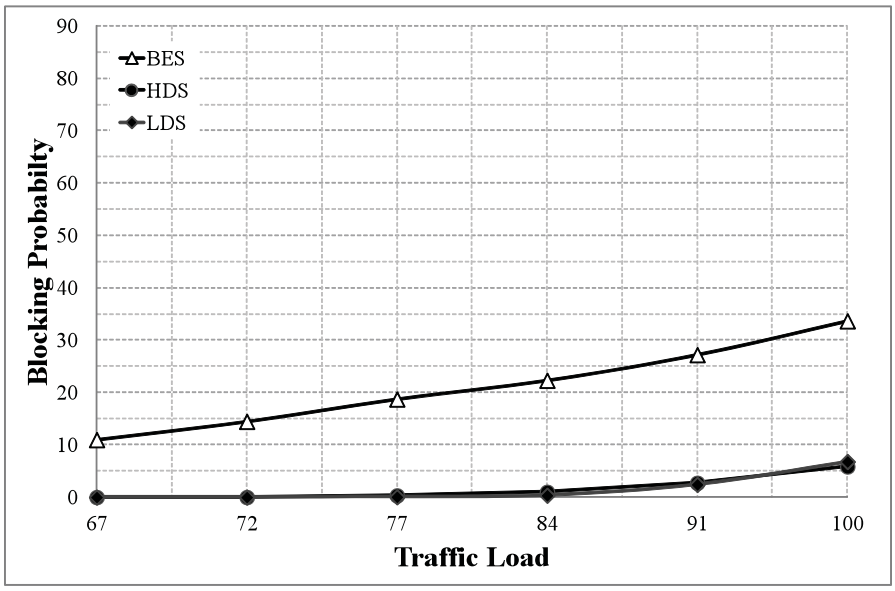

Figure 5.12: Blocking probabilities for the MCRP scheme

5.4 Proposed IaaS Request Admission Control (IRAC) Scheme

In this section, the operation of the proposed IaaS Request Admission Control (IRAC) scheme is described. IRAC module is assumed to be implemented as part of a HCC architecture variant regarding flexible multi-cloud architectures presented in [258]. As already stated in the section 5.2.2, the HCC resources provisioning problem is formulated from the perspective of a single organization, enterprise, institute, etc. assessing that IRAC scheme can make dynamic decisions about migrating workloads to external infrastructure providers (i.e. community and public cloud), when needed. Furthermore, IRAC scheme should be user-centered and flexible enough so as to be easily adapted to various real market applicability cloud infrastructure scenarios such as IHCIs, CSCIs and OCIs described in 5.1.2.

The initial step for the deployment of the IRAC scheme is to define the framework in which it can be applied. In this framework, the various types of user groups and IaaS requests should be identified and classified accordingly. Thus, each user group and each type of IaaS request can be treated differently, allowing the cloud administrator to apply desirable business-level policies, which are mapped to corresponding lower-level IRAC actions (see table 5.4).

Table 5.4: Examples of higher-level policies with corresponding lower-level resource management actions

Business-level policies	IaaS Request Admission Control (IRAC) actions	Scenario with greatest impact
Cloud end users QoS provisioning policy	Ensure continuous availability of resources for privileged virtual desktops	IHCI
Security level VM classification policy	Keep same security-level VMs together, allocate security-sensitive VMs in private partition	All
Power-preservation policy	Consolidation of VMs in the minimum number of physical machines	IHCI
Outsourcing costs minimization policy	Allocate only community-critical IaaS requests to community clouds and all other non-business critical requests to cheaper public clouds	CSCI
Private cloud infrastructure development decision policy	Monitor capacity utilization and QoS-related metrics providing appropriate triggers and alerts	OCI

5.4.1 IaaS requests' and user groups' classification

As depicted in figure 5.13, two main User Groups (UG), namely internal (IUG) and external (EUG) users group are defined. The former group includes trusted users who are members of the assumed organization, enterprise or institute and are registered and thus well-known to the cloud administrator. External users are assumed to be guests and generally untrusted users who temporarily utilize the HCC infrastructure. The definition of these two main user groups is essential for basic QoS provisioning and can be sufficient for simple cloud deployment cases. However, in more demanding scenarios, where there is a requirement to further differentiate

the QoS, users' trust level and access control privileges, each of the main user groups can be divided to two or more subgroups as needed. Furthermore, as the users' behavior is continuously monitored, their initial classification into UGs is periodically adjusted. Therefore, for example, a user may be moved from a high trust UG to a lower trust UG and vice versa. In conclusion, classification is a dynamic procedure, which aims to group together users with same characteristics.

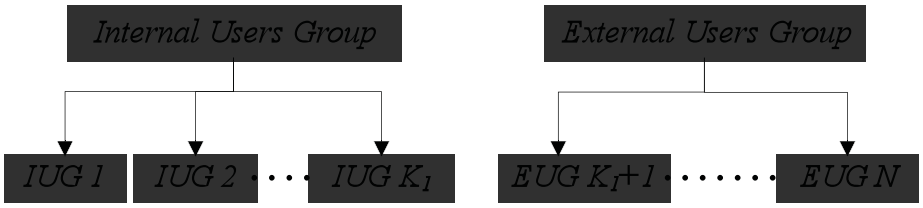

Figure 5.13: Grouping of cloud end users for QoS provisioning purposes

As shown at figure 5.1, each application is related to a single virtual machine (VM) instance and requires a set of CPU, storage and network resources. Consequently, the IaaS requests can be classified based on the requirements of the corresponding applications and thus, those with similar resource requirements are classified to the same Service Group (SG) and can be treated equally by the admission control process, provided that they are initiated by the same UG.

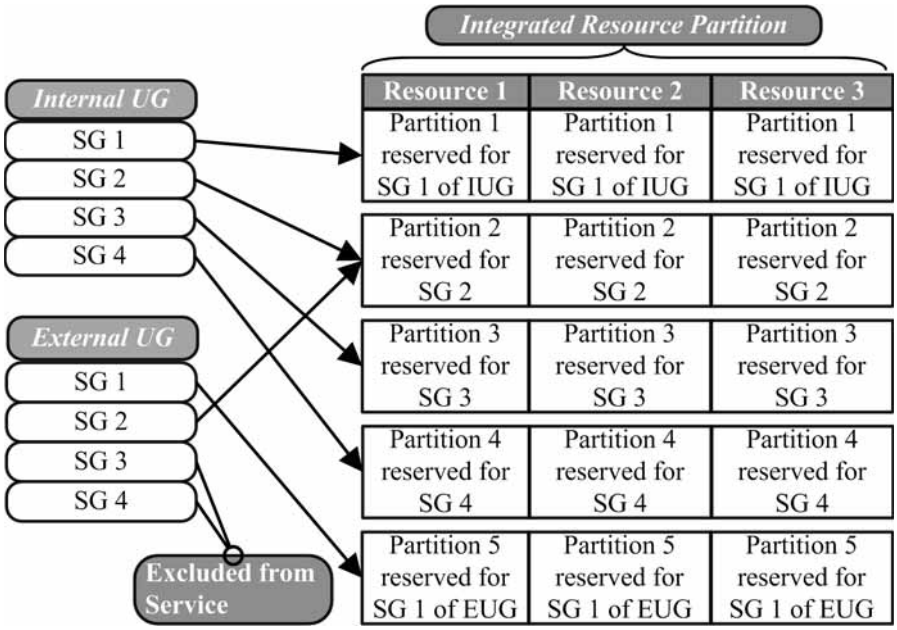

Figure 5.14: Example of mapping the IaaS requests of a user group to partitions

5.4.2 Integrated infrastructure pool partitioning

Resource partitioning is a classic concept with proven efficiency in providing a predetermined acceptance rate for a known load of service requests. However, in the assumed integrated infrastructure pool partitioning case, this concept has to be extended and applied to multiple resources (i.e. CPU, storage, network). Thus, each of the available infrastructure resources are

divided into Z partitions. The number of partitions is determined by the number of UGs' and SGs' combinations (e.g. see at figure 5.14 that SG 2 can be initiated by either IUG or EUG), that require different QoS in terms of resource availability, security etc. The size of each partition is defined so as the expected peak traffic load can be served without exceeding a predefined blocking probability, while any unallocated resources left after the setting of the partitions are equally shared among them (see negotiation phase at figure 5.15). Finally, a set of partitions of different resources is defined for each UG/SG combination. In order to simplify the terminology, in the following, we will refer to this set of resources as Integrated Resource Partition (IRP) as shown at figure 5.14. In other words, all IaaS requests belonging to the same service and user group are mapped to be served by the same integrated partition and thus they experience the same QoS.

An example of IaaS request mapping for the two main user groups is shown in figure 5.14. For the internal users' group, the mapping is straightforward as each SG is mapped to its respective partition. In contrast, for the EUG, two SGs (i.e. SG 3 and 4) are excluded from service and therefore cannot be utilized by the users of the specific group. However, the same users are permitted to access services of SG 2 with the same QoS as the internal users. Furthermore, the users of EUG may also have access to the services of SG 1 utilizing however a different partition than the internal users and having thus different QoS.

In order to confront short-term variations of the traffic load composition, each IRP is allowed to be able to accept non-native IaaS requests (i.e. requests, which were initially aimed to be served by other partitions). However, this has to be performed in a controlled manner in order to prevent the flooding of the IRPs with non-native requests. Hence, for each partition of an IRP two areas are defined, the commonly shared area S and the reserved area B. While a "non-native" service call can be placed only at the commonly shared area, a "native" service call (i.e. a call, which is mapped to be served by the specific IRP) can be placed in any of the two pre-referred areas. The process of allocating capacity to incoming IaaS requests starts from the S area, where native and non-native calls can be accepted. When the commonly shared area becomes fully occupied, then it continues by allocating capacity from the B area, however in this case only to native service calls.

5.4.3 IRAC algorithm

Based on the framework defined in the previous subsections, we will describe in the following the proposed IaaS Request Admission Control algorithm as shown in figure 5.15. Upon the arrival of an IaaS request $R_{i,j}$ that belongs to SG_i and is initiated from user group j, IRAC has to check its partition mapping. In the case that IaaS request is not mapped to an IRP (i.e. excluded from service), it is rejected. Otherwise, IRAC initially checks if the service request can be

accommodated to its native partition. Else, the algorithm sequentially checks if the call can be accommodated to the commonly shared area of the other IRPs.

If this step fails but there are some unallocated resources available, then it is determined, through a negotiation phase, if the user is willing to temporarily use fewer resources than he requires. In this case, the IaaS request is temporarily accepted as "partially satisfied", while an Asynchronous Resource Update (ARU) procedure is responsible to allocate the proper resources to the specific request as soon as they become available.

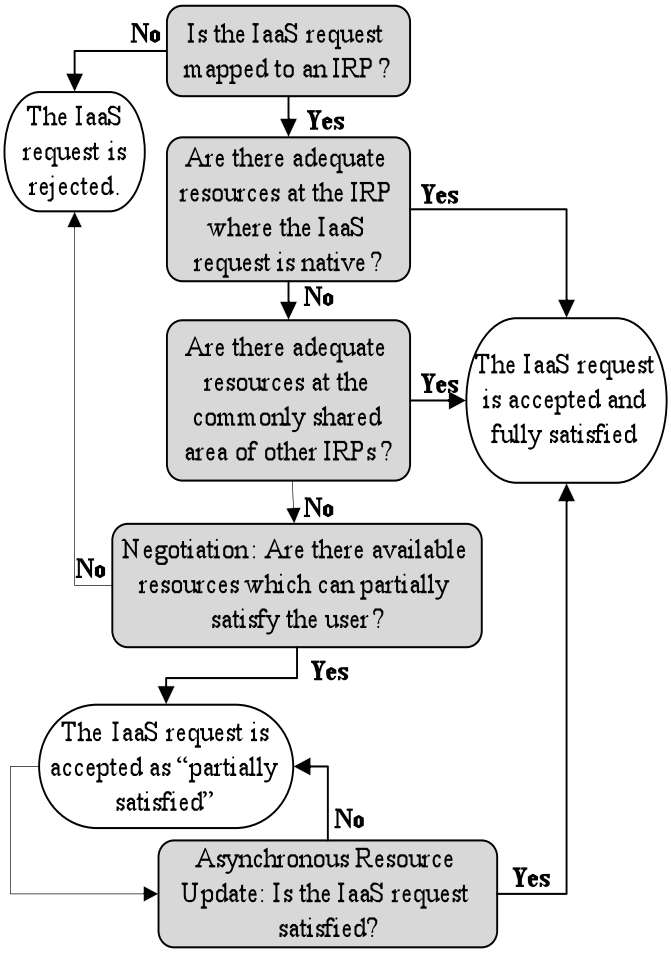

Figure 5.15: Flowchart of the IRAC algorithm

5.4.4 Performance evaluation results

The proposed admission control framework described at the previous section is evaluated through event driven simulation written in C/C++. IaaS requests are assumed to arrive according to a Poisson process, while their duration is exponentially distributed. The performance of IRAC is evaluated in comparison with the Complete Sharing Scheme (CSS), which is a typical resource sharing scheme for CC environments [249]. In CSS, the administrator is able to define UGs and reserve resources for each UG. However, in contrary to

IRAC, the reserved resources for each UG are utilized in the form of a common pool of virtualized computing resources.

As shown in figure 5.14, it is assumed that three main infrastructure resources are shared among the cloud end users. These resources could be CPUs, memory, storage, and other required network by the applications. However, for the sake of generality we will generally refer to them as Resource 1, 2 and 3, assuming that each service requires a number of Basic Resource Units (BRUs) from each of them. Thus, each SG_n has a requirement vector R_n in the form of (R_1 BRU_1, R_2 BRU_2, R_3 BRU_3), while for simplicity of notation R_B=(1,1,1) is set as the basic requirement vector. Regarding users, a simple case of the two main UGs (i.e. IUG and EUG) is assumed as shown in figure 5.13.

In the first simulation scenario, the ability of IRAC to provide QoS provisioning within the same UG is studied. Hence, the focus is on IUG. It is also assumed that users belonging to IUG have access to a low demanding on resources service (LS) of SG 1, a high demanding service (HS) of SG 2 as well as to a best effort service (BF) of SG 3. LS has a basic requirement vector of R_{LS}=R_B, HS has a requirement of R_{HS} =16·R_B and BF service typically requires R_{BF} =5·R_B. We also assume that both HS and LS services require a blocking probability of less than 2%, while the BF service is served as long as there are available resources. The traffic load composition is set to: HS 20%, LS 40 % and BF 40%. In order to make the efficiency of IRAC more obvious, we restrict the available resources assuming a private cloud without any access to community or public cloud resources.

As shown in figure 5.16, the blocking probability is kept (as required) below 2% for both the HS and LS service requests, when the IRAC scheme is employed. At the same time, with CSS the blocking probability of high demanding HS services reaches 30%, while the blocking probabilty of LS services slightly exceeds the 2% threshold at the maximum traffic load. As expected and illustrated in figure 5.17, the blocking probability for BF services is lower with CSS than with IRAC, however this is the trade-off for providing the requested QoS to the HS and LS services and is confronted by IRAC's negotiation and ARU's procedures described at figure 5.15.

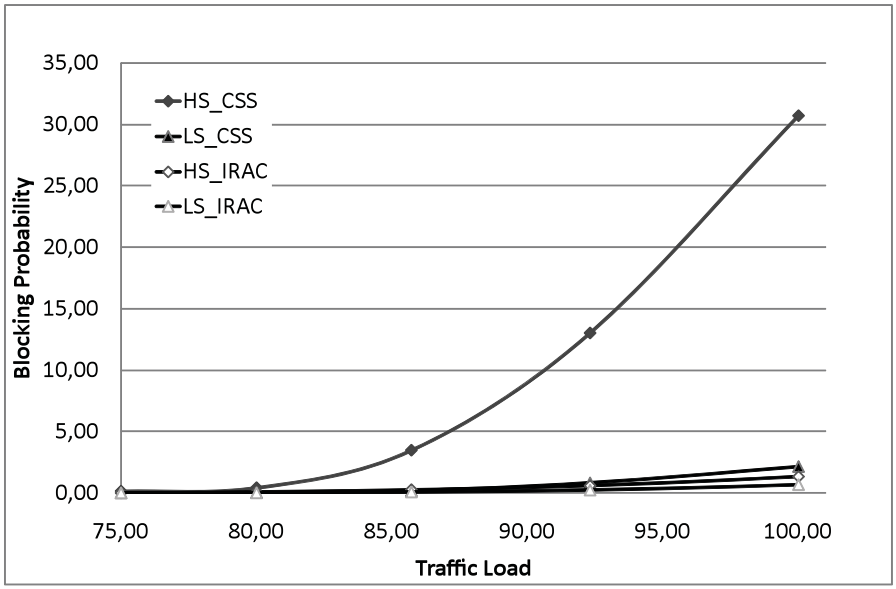

Figure 5.16: Blocking probabilities for HS and LS services in CSS and IRAC schemes

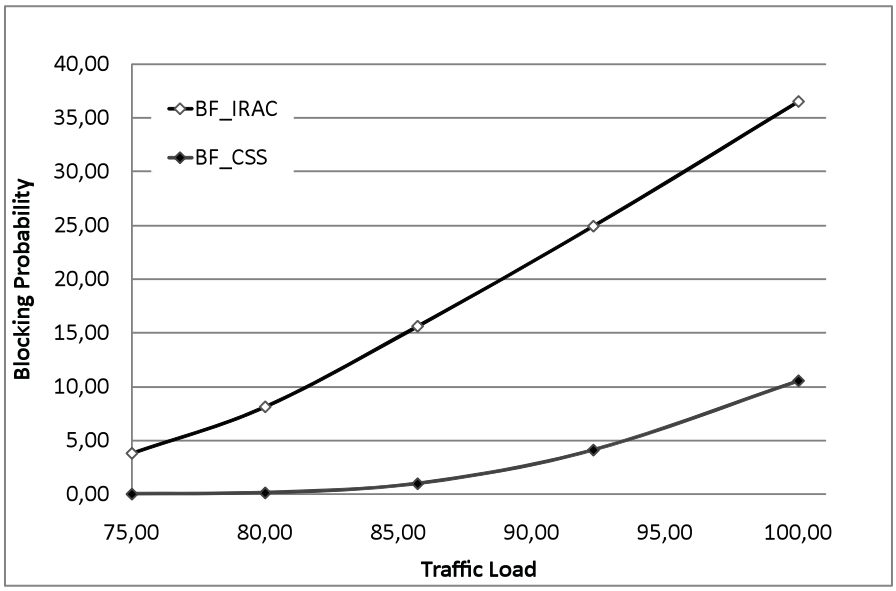

Figure 5.17: Blocking probabilities for BF services in CSS and IRAC schemes

Regarding the second simulation scenario, due to security reasons, IRAC is important to be able to differentiate the way that actual physical resources are allocated to a high security UG in contrast to a lower security UG. In this case, a virtual integrated partition may correspond only to the secured private infrastructure, while other partitions may correspond to resources from the community or public clouds. Thus, typical admission control schemes such as CSS are able to allocate resources for the high security UGs from the private infrastructure, while the outsourced infrastructure resources are allocated to the lower security UGs.

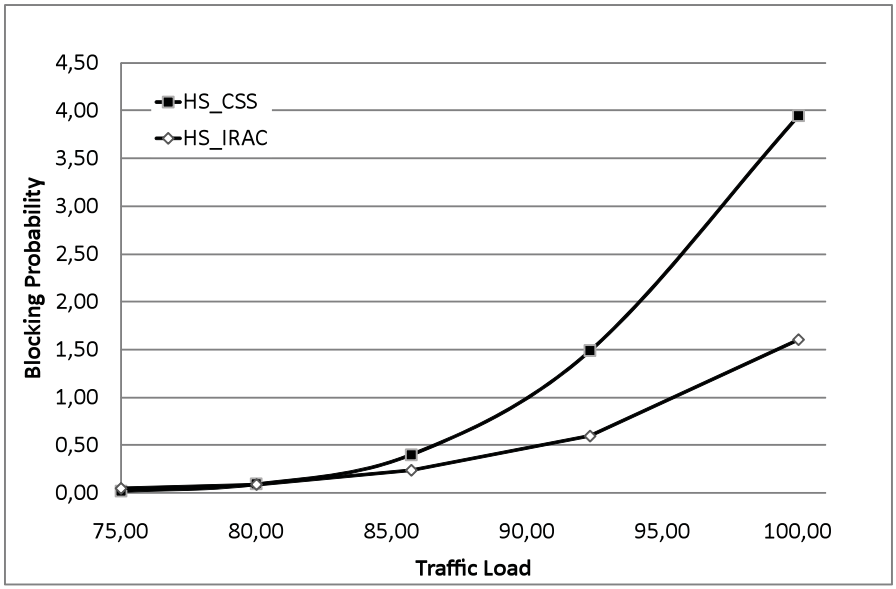

Figure 5.18: Blocking probabilities for HS services in CSS and IRAC schemes

IRAC takes this concept one step further by identifying that different security levels may also be applicable for different kind of services (SGs) initiated from the same UG. Therefore, we extend the first simulation scenario by assuming that IUG is a high security UG while EUG is a lower security UG. The traffic initiated from IUG is 60% of the total traffic load, while the rest 40% comes from EUG. Furthermore, it is defined that only HS of IUG is a security-critical service, while LS of IUG requires lower security levels. Moreover, it is assumed that EUG has access only to BF service, which is considered as inactive for IUG. Consequently, the traffic load composition for EUG is 100% BF services, while for IUG is set to: HS 33% and LS 67 %. In order to demonstrate the efficiency of IRAC in such cases, we further assume that the private infrastructure is adequate to serve only the traffic load of HS services with a blocking probability of less than the required threshold of 2%, while the outsourced infrastructure resources (community or public cloud) are considered as practically unlimited.
Indeed, as shown in figure 5.18, IRAC scheme achieves a blocking probability for the HS service below 2% at the maximum traffic load. This comes as a result of the ability of IRAC to identify that LS services are not security-critical and can be served by outsourced resources. On the other hand, CSS by accepting the LS services at the private cloud almost doubles the blocking probability to approximately 4%.

5.5 Summary

In this chapter, the problem of future networking and computing continuums convergence via the realization of the MCC paradigm has been introduced. The main contribution of the chapter lies in the assertion that in the integrated networking and computing continuum, state-of-the-art resource management frameworks and techniques have to be enhanced in order to

confront the related research challenges from both networking and computing perspectives simultaneously and thus basic design guidelines for the development of corresponding resource management frameworks have been investigated. A typical MCC and HCC environment was assumed as system model for the undertaken research, while MCC/HCC resources provisioning problem was formulated. Two context aware resource management schemes for mobile/hybrid cloud infrastructures have been proposed, namely MCRP and IRAC. Performance evaluation results have shown that being aware of both radio access and computing resources of a customized MCC/HCC infrastructure can enhance state-of-the-art resource management frameworks in terms of QoS provisioning and overall system capacity utilization by taking advantage of CC technology's strengths. Conclusively, a user-oriented and customizable infrastructure sharing approach can be adopted in order efficient mobile cloud services provisioning solutions to be realized.

CHAPTER 6

CONCLUSIONS & FUTURE WORK

6.1 Thesis Summary

In this PhD thesis, the research problem called context aware resource management for mobile and fixed networking systems has been investigated. In the introductory chapter, the research motivation, scope, objectives and anticipated impact have been discussed. More specifically, the notion of context, context information and context awareness for mobile and fixed networking systems convergence have been defined in order the reader to have a clear view of this thesis' technical contributions from the very beginning. The term "context aware resource management" has also been defined in order to outline the innovative features that are introduced in this thesis in comparison with state-of-the-art architectural and algorithmic resource management solutions being available in the international research community nowadays. Chapter 2 includes all literature review and state-of-the-art related works. Hence, a novel research field with promising expectations for further future research called context aware mobile and wireless networking (CAMoWiN) has been surveyed, while various context aware functionalities have been thoroughly analyzed. State-of-the-art context aware resource management algorithms and schemes have also been surveyed and hints about possible contributions and open issues that the current thesis will deal with have been given. Chapters 3, 4 and 5 include the purely technical contributions of the current thesis. According to the three pillars of the thesis' roadmap depicted in figure 1.1, each chapter's content is dedicated to context aware resource management architectural and algorithmic innovations from the perspective of one of these three pillars at a time, namely: a) 4G Hetnet environment (chapter 3), b) mobile and fixed networking systems' convergence (chapter 4), and c) hybrid/mobile cloud infrastructure (chapter 5). In the following subsections, the basic results and contributions for the pre-mentioned research pillars are summarized.

6.1.1 4G HetNet environment

In chapter 3, the first research pillar of the current thesis has been investigated dealing with context aware resource management aspects being applicable in a 4G HetNet environment. After defining the research problem and challenges regarding 4G small cell networks, a novel concept called femto-relaying has been proposed, which aims to take advantage of both femto/small cell and wireless relaying technologies. Two context aware frameworks for the efficient integration of small cells in the existing IP and cellular infrastructures have been proposed and numerous performance evaluation results have been conducted to make proof-of-concept considerations of the proposed femto-relay concept. More specifically, simulation results have also shown that the overall performance of a HetNet environment can be

leveraged in terms of QoS provisioning for various MT/UE groups and Hetnet deployment settings, energy saving, data rate/capacity and area spectral efficiency enhancements.

6.1.2 Mobile and fixed networking systems' convergence

Chapter 4 addresses the second research pillar of the current PhD thesis. Here, the novel concept of the integrated services router (ISR) has been introduced accompanied by real/future realistic market network deployment scenarios being applicable and interrelated with a 4G Hetnet environment. Three context aware resource management schemes have been proposed residing at novel and emerging real-market products called small cell and machine type communication gateways. The proposed resource management modules take simultaneously into consideration both available radio and backhaul resources in order to realize efficient decision making procedures in terms of QoS provisioning and overall system capacity/utilization. Hence, it is shown that a converged 4G HetNet deployment setting consisting of both mobile and fixed networking subsystems can better deal with emerging and up-to-date resource management problems.

6.1.3 Mobile Cloud Computing/Networking

Chapter 5 deals with the third research pillar of the current thesis (i.e. hybrid cloud infrastructure). For mobile/hybrid cloud computing infrastructures, novel context aware resource management schemes are proposed, while the role that cloud computing (CC) technology can play in the evolution of context aware mobile and wireless networking research (CAMoWiN) is the main field of the study. Hence, typical mobile cloud computing (MCC) and hybrid cloud computing environments (HCC) are considered, while the research problem is formulated based on the need for context aware MCC/HCC resources management. More specifically, two novel schemes have been proposed, that is MCRP for mobile cloud resources provisioning and IRAC for HCC resources provisioning. The performance evaluation results provided a proof-of-concept analysis of the fact that in the integrated networking and computing continuum, state-of-the-art resource management frameworks and techniques have to be enhanced in order to confront the related research challenges from both networking and computing perspectives simultaneously. In other words, it has been shown that being aware of both radio access and computing resources of a customized MCC/HCC infrastructure can enhance state-of-the-art resource management frameworks in terms of QoS provisioning and overall system capacity utilization by taking advantage of CC technology's strengths. Conclusively, a user-oriented and customizable infrastructure sharing approach can be adopted in order efficient mobile cloud services provisioning solutions to be realized.

6.2 Future Research Directions

Hints for future research directions have already been discussed in individual parts of the current PhD thesis. The novel notion of context aware resource management is expected to be applicable to many research variants in the beyond 4G era. Hence, in the 5G era, the so called trend of cells' densification will continue to emerge aiming to exploit these shrunken cells in order to put massive amounts of bandwidth precisely where people or machines are using it (e.g. malls, arenas, public plazas, urban parks, busy business districts, etc). Backhaul challenges will surely be a major research problem in terms of traversing the large amounts of mobile traffic via both wireless and wired networking infrastructures. Wireless backhaul alternatives are expected to be more in some years from now (e.g. via usage of millimetre wave frequencies of 60/90 GHz), providing thus more degrees of freedom regarding the best alternative route in order specific QoS, energy, security, overall system capacity etc requirements to be met. 5G technologies are also expected to provide: a) higher system spectral efficiency, b) lower energy consumption, c) lower outage probabilities, d) higher bit rates in larger portions of the coverage areas, e) lower latencies, f) higher number of supported devices, g) lower infrastructure deployment costs, h) higher versatility, scalability and enhanced reliability. Integration of MTC devices to enable the Internet of Things with vast numbers of connected devices as well as novel applications such as mission critical control, traffic safety, smart grid and many others remains a great challenge.

Regarding seamless integration of CC technology strengths at the application layer and mobile wireless networks at transmission and network layers, even more enhanced context aware functionalities will be required in order end-to-end mobile cloud services to be efficiently delivered (i.e. mobile cloud service = mobile network + computing + storage). The integration of the successful virtualization model and its applicability in mobile and wireless networking will ultimately provide innovative and flexible architectural innovations in which MCC/MCN assets (e.g. network, processing power, storage resources, etc) could move around and be migrated whenever/wherever it is needed in the 5G network infrastructure.

References

[1] W. N. Schilit, ”A system architecture for context-aware mobile computing,” PhD Thesis, Columbia University, New York, 1995.

[2] N. Malik, U. Mahmud and Y. Javed, “Future Challenges in Context- Aware Computing”, IADIS International Conference WWW/Internet, 2007.

[3] A. A. Lazar, and G. Pacifici, “Control of resources in broadband networks with quality of service guarantees”, IEEE Communications Magazine, vol. 29(10), pp. 66-73, October 1991.

[4] D. Ferrari, D. C. Verma, “A scheme for real-time channel establishment in wide-area networks”, IEEE Journal on Selected Areas in Communications, vol. 8(3), pp. 368-379, April 1990.

[5] D. D. Clark, S. Schenker and L. Zhang, **“**Supporting real-time applications in an Integrated Services Packet Network: architecture and mechanism”, ACM SIGCOMM Computer Communication Review, vol. 22(4), pp. 14-26, October 1992.

[6] G. De Veciana, G. Kesidis and J. Walrand, “Resource management in wide-area ATM networks using effective bandwidths”, IEEE Journal on Selected Areas in Communications, vol. 13(6), pp. 1081-1090, August 1995.

[7] P. Bahl, I. Chlamtac and R. Farago, “Resource assignment for integrated services in wireless ATM networks”, International Journal of Communication Systems, vol. 11, pp. 29-41, 1998.

[8] S. C. Borst and D. Mitra, “Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic”, IEEE Journal on Selected Areas in Communications, vol. 16(5), pp. 668-678, June 1998.

[9] I. Hsu and J. Walrand, “Admission control and resource management for multi-service ATM networks”, Springer Telecommunication Systems Journal, vol. 7(1-3), pp. 185-207, June 1997.

[10] S. K. Das and C. Rose, ”Coping with Uncertainty in Mobile Wireless Networks”, in Proc. 15th IEEE International Symposium on Personal Indoor and Mobile Radio Communications, pp. 103-108, PIMRC 2004.

[11] M. A. Razzaque, S. Dobson and P. Nixon, ”Cross-Layer Architectures for Autonomic Communications”, Springer Network and Systems Management Journal, vol. 15(1), pp. 13-27, 2007.

[12] E. Gustafsson and A. Jonsson, “Always best connected”, IEEE Wireless Communications Magazine, vol. 10(1), pp. 49-55, February 2003.

[13] B. Jennings, S. V. Meer, S. Balasubramaniam, D. Botvich, M. O’Foghlu, W. Donnelly and J. Strassner, “Towards Autonomic Management of Communications Networks”, IEEE Communications Magazine, vol. 45(10), pp. 112-121, 2007.

[14] S. Dobson, S. Denazis, A. Fernandez, D. Gaiti, E. Gelenbe, F. Massacci, P. Nixon, F. Saffre, N. Schmidt and F. Zambonelli, “A Survey of Autonomic Communications”, ACM Transactions on Autonomous and Adaptive Systems, vol. 1(2), pp. 223-259, 2006.

[15] N. Agoulmine, S. Balasubramaniam, D. Botvitch, J. Strassner, E. Lehtihet and W. Donelly, “Challenges for Autonomic Network Management”, proceedings of the 1st conference on Modelling Autonomic Communication Environment (MACE), Ireland, 2006.

[16] J. Belschner et al., “Optimization of Radio Access Network Operation Introducing Self-x Functions: Use Cases, Algorithms, Expected Efficiency Gains”, proceedings of IEEE 69th Vehicular Technology Conference, VTC Spring 2009.

[17] 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(3), pp. 337-357, 2011.

[18] J. Strassner, S. V. Meer, D. O’Sullivan and S. Dobson, “The Use of Context-Aware Policies and Ontologies to Facilitate Business-Aware Network Management”, Journal of Network and Systems Management, vol. 17(3), pp. 255-284, 2009.

[19] K. Wrona and L. Gomez, “Context-Aware Security and Secure Context- Awareness in Ubiquitous Computing Environments”, proceedings of 21st Conference of Polish Information Processing Society (PIPS), pp. 255- 265, 2005.

[20] X. Jiang and J. A. Landay, “Modelling Privacy Control in Context-Aware Systems”, IEEE Pervasive Computing, vol. 1(3), pp. 59-63, 2002.

[21] J. Hong, E. H. Suh, J. Kim and S. Kim, “Context-Aware System for Proactive Personalized Service based on Context History”, Elsevier Expert Systems with Applications Journal, vol. 36(4), pp. 7448-7457, 2009.

[22] H. J. La and S. D. Kim, “A Conceptual Framework for Provisioning Context-Aware Mobile Cloud Services”, proceedings of IEEE 3rd International Conference on Cloud Computing, pp. 466-473, 2010.

[23] 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(1), pp. 362-386, 2013.

[24] P. Makris, D. N. Skoutas, C. Skianis, “On networking and computing environments' integration: A novel mobile cloud resources provisioning approach”, proceedings of IEEE International Conference on Telecommunications and Multimedia (TEMU), pp. 71-76, 2012.

[25] M. Vrdoljak, S. I. Vrdoljak, G. Skugor, “Fixed-mobile convergence strategy: technologies and market opportunities”, IEEE Communications Magazine, vol. 38(2), pp. 116-121, February 2000.

[26] Dong-Hoon Yang, Seongcheol Kim; Changi Nam; Ji-Sook Moon, “Fixed and mobile service convergence and reconfiguration of telecommunications value chains”, IEEE Wireless Communications, vol. 11(5), pp. 42-47, October 2004.

[27] P. Makris, N. Nomikos, D. N. Skoutas, D. Vouyioukas, C. Skianis, J. Zhang and C. Verikoukis, “A Context Aware Framework for the Efficient Integration of Femtocells in IP and Cellular Infrastructures”, EURASIP Journal on Wireless Communications and Networking, Special Issue on Small Cell Cooperative Communications, 2013:62, available online at http://jwcn.eurasipjournals.com/content/2013/1/62.

[28] M. Dohler, T. Watteyne, and J. Alonso, “Machine-to- Machine: An Emerging Communication Paradigm,” Tutorial, PIMRC 2010, 26 Sept. 2010, Istanbul, Turkey; also at Globecom 2010, Miami 2010, USA.

[29] P. Makris, D. N. Skoutas, N. Nomikos, D. Vouyioukas and C. Skianis, "A Context-Aware Backhaul Management Solution for combined H2H and M2M traffic", proceedings of IEEE International Conference on Computer, Information and Telecommunication Systems (CITS 2013), 7-8 May, Piraeus, Greece.

[30] D. Lopez-Perez, I. Guvenc, G. De La Roche, M. Kountouris, T. Q. S. Quek and J. Zhang, "Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks", IEEE Wireless Communications, vol. 18(3), pp. 22-30, 2011.

[31] J. Zhang and G. De La Roche, "Femtocells: Technologies and Deployment" John Wiley & Sons Ltd, ISBN 978-0470742983, 2010.

[32] D. N. Skoutas, P. Makris and C. Skianis, "Optimized Admission Control Scheme for Coexisting Femtocell, Wireless and Wireline Networks", Springer Telecommunication Systems Journal, Special Issue on Mobility Management in Future Internet, in press, 2013.

[33] C. Skianis, "Radio vs. Backhaul Bottlenecks: An Integrated Quality of Service Provisioning Approach for Small Cell Gateways", International Journal of Communication Systems, available online at http://onlinelibrary.wiley.com/doi/10.1002/dac.2541/abstract

[34] P. Makris, D. N. Skoutas, P. Rizomiliotis and C. Skianis, "A User-Oriented, Customizable Infrastructure Sharing Approach for Hybrid Cloud Computing Environments", proceedings of 3rd IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pp. 432-439, 29/11-01/12, Athens, Greece, 2011.

[35] P. J. Brown, J. D. Bovey and X. Chen, "Context-aware applications: From the laboratory to the marketplace", IEEE Personal Communications, vol. 4(5), pp. 58-64, 1997.

[36] R. Hull, P. Neaves and J. Bedford-Roberts, "Towards situated computing", proceedings of 1st International Symposium on Wearable Computers (ISWC), pp. 146-153, Cambridge, 1997.

[37] D. Franklin and J. Flaschbart, "All gadget and no representation makes jack a dull environment", proceedings of the AAAI Spring Symposium on Intelligent Environments, pp. 155-160, Palo Alto, 1998.

[38] A. K. Dey, "Understanding and Using Context", Personal and Ubiquitous Computing Journal, vol. 5(1), pp. 4-7, 2001.

[39] J. Strassner and D. O'Sullivan, "Knowledge Management for Context-Aware Policy-based Ubiquitous Computing Systems", proceedings of 6th International Workshop on Managing Ubiquitous Communications and Services, pp. 67-76, Spain, 2009.

[40] K. Henricksen, "A framework for context-aware pervasive computing applications", PhD thesis, School of Information Technology and Electrical Engineering, The University of Queensland, Australia, 2003.

[41] R. Sterritt, M. Mulvenna and A. Lawrynowicz, "Dynamic and Contextualised Behavioural Knowledge in Autonomic Communications", Springer Verlag Berlin Heidelberg, LCNS 3457, pp. 217-228, 2005.

[42] John Strassner et al., "The Use of Context-Aware Policies and Ontologies to Facilitate Business-Aware Network Management", Journal of Network and Systems Management, vol. 17(3), pp. 255-284, 2009.

[43] M. Charalambides et al., “Policy Conflict Analysis for DiffServ Quality of Service Management”, IEEE Transactions on Network and Service Management, vol. 6(1), pp. 15-30, 2009.

[44] P. D. Reyes, J. Favela and J. C. Castillo, “Uncertainty Management in Context-Aware Applications: Increasing Usability and User Trust”, Springer Wireless Personal Communications Journal, vol. 56(1), pp. 37-53, 2011.

[45] K. Henricksen and J. Indulska, “Modelling and Using Imperfect Context Information”, proceedings of the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW), pp. 33-37, 2004.

[46] R. Schmohl and U. Baumgarten, “Context-aware Computing: a Survey Preparing a Generalized Approach”, proceedings of the International Multi-Conference of Engineers and Computer Scientists, Hong Kong, 2008.

[47] K. Henricksen et al., “Middleware for Distributed Context-Aware Systems”, Springer Verlag Berlin Heidelberg, LNCS 3760, pp. 846-863, 2005.

[48] R. Winter et al., “CrossTalk: Cross-Layer Decision Support based on Global Knowledge”, IEEE Communications Magazine, vol. 44(1), pp. 93-99, 2006.

[49] L. Sarakis, G. Kormentzas and F. M. Guirao, “Seamless Service Provision for Multi Heterogeneous Access”, IEEE Wireless Communications, vol. 16(5), pp. 32-40, 2009.

[50] G. Chen and D. Kotz, “A Survey of Context-Aware Mobile Computing Research”, Dartmouth Computer Science Technical Report TR2000-381, Hanover, 2000.

[51] E. Christopoulou, C. Goumopoulos and A. Kameas, “An Ontology-Based Context Management and Reasoning Process for UbiComp Applications”, proceedings of Joint sOc-EUSAI'2005 Conference, pp. 265-270, ACM Press, Grenoble, France, 2005.

[52] G. D. Abowd et al., “Cyberguide: A Mobile Context-Aware Tour Guide”, Baltzer/ACM Wireless Networks Journal, vol. 3(5), pp. 421-433, 1997.

[53] N. Davies et al., “Limbo: A Tuple Space Based Platform for Adaptive Mobile Applications”, proceedings of the International Conference on Open Distributed Processing/Distributed Platforms (ICODP/ICDP), pp. 291-302, Toronto, Canada, 1997.

[54] A. Schmidt et al., “Advanced interaction in context”, proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (HUC), pp. 89-101, Karlsruhe, Germany, 1999.

[55] H. W. Gellersen, A. Schmidt and M. Beigl, “Multi-Sensor Context-Awareness in Mobile Devices and Smart Artifacts”, Mobile Networks and Applications (MONET) Journal, vol. 7(5), pp. 341-351, 2002.

[56] P. Gray and D. Salber, “Modelling and Using Sensed Context Information in the Design of Interactive Applications”, 8th IFIP International Conference on Engineering for Human-Computer Interaction, LNCS 2254, pp. 317-336, Springer, 2001.

[57] G. Chen and D. Kotz, “Solar: A Pervasive-Computing Infrastructure for Context-Aware Mobile Applications”, Technical Report TR2002-421, Dartmouth College, 2002.

[58] C. Anagnostopoulos, A. Tsounis and S. Hadjiefthymiades, "Context Awareness in Mobile Computing Environments: a Survey", Springer Wireless Personal Communications Journal, vol. 42(3), pp. 445-464, 2007.

[59] K. E. Kjær, "A Survey of Context-Aware Middleware", proceedings of 25th conference on International Multi-Conference: Software Engineering Innsbruck, pp. 148-155, Austria, 2007.

[60] P. Debaty et al., "Integrating the Physical World with the Web to Enable Context-Enhanced Services", Springer Mobile Networks and Applications Journal, vol. 10(4), pp. 385-394, 2005.

[61] H. Chen, T. Finin and A. Joshi, "An Ontology for Context-Aware Pervasive Computing Environments" Knowledge Engineering Review, Special Issue on Ontologies for Distributed Systems, vol. 18(3), pp. 197–207, 2004.

[62] T. Gu, H. K. Pung and A. Joshi, "A Service-Oriented Middleware for Building Context-Aware Services" Elsevier Journal of Network and Computer Applications (JNCA), vol. 28(1), pp. 1–18, 2005.

[63] W. Qin, Y. Suo and Y. Shi, "CAMPS: A Middleware for Providing Context-Aware Services for Smart Space" proceedings of Advances in Grid and Pervasive Computing, Springer Verlag Berlin Heidelberg, LNCS 3947, pp. 644–653, 2006.

[64] H. V. Kranenburg et al., "A Context Management Framework for Supporting Context-Aware Distributed Applications", IEEE Communications Magazine, vol. 44(8), pp. 67-74, 2006.

[65] P. Fahy and S. Clarke, "CASS: Middleware for Mobile, Context-Aware Applications", Workshop on Context Awareness at MobiSys '04, Boston, 2004.

[66] A. Chan and S. N. Chuang, "MobiPADS: A Reflective Middleware for Context-Aware Mobile Computing", IEEE Transactions on Software Engineering, vol. 29(12), pp. 1072 – 1085, 2003.

[67] The CONTEXT project, http://context.upc.es/index.htm, accessed in May 2013.

[68] J. Strassner, "Policy-Based Network Management", Morgan Kaufman Publishers, ISBN 1-55860-859-1, September 2003.

[69] U. Mahmud et al., "Context-Aware Paradigm for a Pervasive Computing Environment (CAPP)", proceedings of IADIS International Conference on WWW/Internet, Villa Real, Portugal, pp. 337-346, 2007.

[70] R. Kodikara, C. Ahlund and A. Zaslavsky, "Towards Context Aware Adaptation in Wireless Networks", proceedings of 2nd International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM), pp. 245-250, 2008.

[71] C. Casetti et al., "Autonomic Interface Selection for Mobile Wireless Users", IEEE Transactions on Vehicular Technology, vol. 57(6), pp. 3666-3678, 2008.

[72] N. Saxena, A. Roy and J. Shin, "CARP: Context-Aware Resource Provisioning for Multimedia over 4G Wireless Networks", proceedings of ICCS, Springer Verlag Berlin Heidelberg, LNCS 4487, pp. 652-659, 2007.

[73] A. Hasswa, N. Nasser and H. Hassanein, "A seamless context-aware architecture for fourth generation wireless networks", Springer Wireless Personal Communications Journal, Special Issue on Seamless Handover in next Generation Wireless/Mobile Networks, vol. 43(3), pp. 1035-1049, 2007.

[74] A. De La Oliva et al., “An Overview of IEEE 802.21: Media-Independent Handover Services”, IEEE Wireless Communications, vol. 15(4), pp. 96-103, 2008.

[75] S. Buljore et al., “Architecture and Enablers for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks: The IEEE 1900.4 Working Group”, IEEE Communications Magazine, vol. 47(1), pp. 122-129, 2009.

[76] S. Fernandes and A. Karmouch, “Vertical Mobility Management Architectures in Wireless Networks: A Comprehensive Survey and Future Directions”, IEEE Communications Surveys & Tutorials, vol. 14(1), pp. 45-63, 2012.

[77] G. Lampropoulos et al., “Enhanced Media Independent Handover Procedure for Next Generation Networks”, proceedings of Future Network and Mobile Summit 2010 Conference, pp. 1-8, Florence, Italy, 2010.

[78] Y. Wang et al., “Handover Management in Enhanced MIH Framework for Heterogeneous Wireless Networks Environment”, Springer Wireless Personal Communications, vol. 52(3), pp. 615-636, 2010.

[79] A. Galani et al., “Design and Assessment of Functional Architecture for Optimized Spectrum and Radio Resource Management in Heterogeneous Wireless Networks”, Wiley International Journal of Network Management, vol. 20(4), pp. 219–241, 2010.

[80] A. I. Wang and Q. K. Ahmad, “CAMF - Context-Aware Machine learning Framework for Android”, International Conference on Software Engineering and Applications (SEA), Marina Del Rey, CA, USA, 2010.

[81] N. Baker et al., “Context-Aware Systems and Implications for Future Internet”, proceedings of Future Internet Assembly Book named “Towards the Future Internet – A European Reasearch Perspective”, edited by G. Tselentis et al., pp. 335-344, 2009.

[82] L. Sliman, F. Biennier and Y. Badr, “A Security Policy Framework for Context-Aware and User Preferences in E-Services”, Elsevier Journal of Systems Architecture, Special Issue on Secure SOA, vol. 55(4), pp. 275-288, 2009.

[83] S. Reddy, J. Burke, D. Estrin, M. Housen and M. Srivastava, “Using Mobile Phones to Determine Transportation Modes”, ACM Transactions on Sensor Networks”, vol. 6(2), pp. 1-27, 2010.

[84] K. Cho, I. Hwang, S. Kang, B. Kim, L. Jinwon, L. Sangjeong, P. Souneil, S. Junehwa and R. Yunseok, “HiCon: A Hierarchical Context Monitoring and Composition Framework for Next-Generation Context-Aware Services”, IEEE Network Magazine, vol. 22(4), pp. 34-42, 2008.

[85] N. Samaan and A. Karmouch, “Towards Autonomic Network Management: an Analysis of Current and Future Research Directions”, IEEE Communications Surveys and Tutorials, vol. 11(3), pp. 22-36, 2009.

[86] S. J. Yoo and N. Golmie, “Policy-based scanning with QoS support for seamless handovers in wireless networks”, Wireless Communications and Mobile Computing Journal, vol. 10(3), pp. 405-425, 2010.

[87] E. Linehan, S. L. Tsang and S. Clarke, “Supporting Context-Awareness: A Taxonomic Review”, TCD-CS-2008-37, October 2008.

[88] S. Kang, L. Jinwon, J. Hyukjae, L. Youngki, P. Souneil and S. Junehwa, “A Scalable and Energy-Efficient Context Monitoring Framework for Mobile Personal Sensor Networks”, IEEE Transactions on Mobile Computing, vol. 9(5), pp. 686-702, 2010.

[89] P. Nurmi, M. Martin and J. A. Flanagan, “Enabling Proactiveness Through Context Prediction”, Workshop on Context Awareness for Proactive Systems (CAPS), Helsinki University Press, pp. 159–168, 2005.

[90] T. Anagnostopoulos, C. Anagnostopoulos, S. Hadjiefthymiades, M. Kyriakakos and A. Kalousis, “Predicting the Location of Mobile Users: A Machine Learning Approach”, ACM International Conference on Pervasive Services (ICPS), Imperial College, London, 2009.

[91] S. Sigg, S. Haseloff, and K. David, “An Alignment Approach for Context Prediction Tasks in UbiComp Environments”, IEEE Pervasive Computing, vol. 9(4), pp. 90-97, 2010.

[92] R. Mayrhofer, H. Radi and A. Ferscha, “Recognizing and Predicting Context by Learning from User Behavior”, Radiomatics: Journal of Communication Engineering, Special Issue on Advances in Mobile Multimedia, vol. 1(1), pp. 30-42, 2004.

[93] Y. Wang, M. Martonosi and L. S. Peh, “Predicting Link Quality using Supervised Learning in Wireless Sensor Networks”, Mobile Computing and Communications Review (MC2R), July 2007.

[94] A. Krause, A. Smailagic and D. P. Siewiorek, “Context-Aware Mobile Computing: Learning Context- Dependent Personal Preferences from a Wearable Sensor Array”, IEEE Transactions on Mobile Computing, vol. 5(2), pp. 113- 127, 2006.

[95] X. Zhu, “Semi-supervised learning”, In Claude Sammut and Geoffrey Webb, editors, Encyclopedia of Machine Learning, Springer, pp. 892-897, 2010.

[96] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan and D. Riboni, “A survey of context modelling and reasoning techniques”, Elsevier Pervasive and Mobile Computing Journal, vol. 6(2), pp. 161-180, 2010.

[97] B. Moltchanov, C. Mannweiler and J. Simoes, “Context-Awareness Enabling New Business Models in Smart Spaces”, Springer Verlag Berlin Heidelberg, LNCS 6294, pp. 13-25, 2010.

[98] B. Siljee, I. Bosloper and J. Nijhuis, “A Classification Framework for the Storage and Retrieval of Context”, Workshop on Modelling and Retrieval of Context (MRC), Ulm, Germany, 2004.

[99] J. Coutaz, J. L. Crowley, S. Dobson and D. Garlan, “Context is Key”, ACM Communications Magazine, vol. 48(3), pp. 49-53, 2005.

[100] D. Balakrishnan, M. E. Barachi, A. Karmouch and R. Glitho, “Challenges in Modeling and Disseminating Context Information in Ambient Networks”, Springer Verlag Berlin Heidelberg, LNCS 3744, pp. 32-42, 2005.

[101] X. Hu et al., “A Hybrid Peer-to-Peer Solution for Context Distribution in Mobile and Ubiquitous Environments”, Springer Information Systems Development, Paphos, Cyprus, pp. 501–510, 2009.

[102] K. Stanoevska-Slabeva, T. Wozniak, I. Hoffend, C. Mannweiler and H. D. Schotten, “The Emerging Ecosystem for Context Information and the Role of Telecom Operators”, 14th International Conference on Intelligence in Next Generation Networks (ICIN), pp. 1-6, 2010.

[103] S. L. Kiani, M. Knappmeyery, N. Baker and B. Moltchanov, "A Federated Broker Architecture for Large Scale Context Dissemination", 10th IEEE International Conference on Computer and Information Technology (CIT 2010), pp. 2964-2969, Bradford, UK, 2010.

[104] A. Klein, C. Mannweiler, J. Schneider and H. D. Schotten, "Access Schemes for Mobile Cloud Computing", 11th International Conference on Mobile Data Management (MDM), pp. 387-392, May 2010.

[105] J. Simões and T. Magedanz, "Contextualized User-Centric Multimedia Delivery System for Next Generation Networks", Springer Telecommunication Systems Journal, vol. 48(3-4), pp. 301-316, 2011.

[106] J. Antoniou, F. C. Pinto, J. Simoes and A. Pitsillides, "Supporting Context-Aware Multiparty Sessions in Heterogeneous Mobile Networks", Elsevier Mobile Networks and Applications Journal, vol. 15(6), pp. 831-844, 2010.

[107] J. Antoniou, C. Christophorou, J. Simoes and A. Pitsillides, "Adaptive Network-Aided Session Support in Context-Aware Converged Mobile Networks", International Journal of Autonomous and Adaptive Communication Systems, vol. 5(3), pp. 201-232, 2012.

[108] A. U. H. Yasar, Y. Vanrompay, D. Preuveneers and Y. Berbers, "Optimizing Information Dissemination in Large Scale Mobile Peer-to-Peer Networks Using Context-based Grouping", 13th International Conference on Intelligent Transportation Systems, pp. 1065-1071, 2010.

[109] P. Makris, G. Lampropoulos, D. Skoutas and C. Skianis "New Directions and Challenges for MIH Operation Towards Real Market Applicability", 1st International Workshop on Mobility in Future Internet (MiFI '10), Crete Greece, 2010.

[110] N. Dimitriou, L. Sarakis, D. Loukatos, G. Kormentzas and C. Skianis, "Vertical Handover Framework for Future Collaborative Wireless Networks", Wiley International Journal of Network Management, vol. 21(6), pp. 548-564, 2011.

[111] IEEE Std 802.21-2008, (2009), "IEEE Standard for Local and Metropolitan Area Networks-Part 21: Media Independent Handover Services", IEEE.

[112] IEEE P1900.4/D1.5 (2008), "Draft Standard for Architectural Building Blocks Enabling Network-device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks", IEEE.

[113] P. Demestichas, "Introducing Cognitive Systems in the Wireless B3G World: Motivations and Basic Engineering Challenges", Elsevier Telematics and Informatics, vol. 27(3), pp. 256-268, 2010.

[114] A. Saatsakis, K. Tsagkaris and P. Demestichas, **"**Exploiting Context, Profiles and Policies in Dynamic Sub-carrier Assignment Algorithms for Efficient Radio Resource Management in OFDMA Networks", Springer Annals of Telecommunications, vol. 65(7), pp. 359-374, 2010.

[115] K. Nolte et al., "The E3 architecture: enabling future cellular networks with cognitive and self-x capabilities", Wiley International Journal of Network Management, vol. 21(5), pp. 360-383, 2011.

[116] H. Derbel, N. Agoulmine and M. Salaün, “ANEMA: Autonomic Network Management Architecture to Support Self-Configuration and Self-Optimization in IP Networks, Elsevier Computer Networks, vol. 53(3), pp. 418–430, 2009.

[117] P. Makris and C. Skianis, “Multi-Scenario Based Call Admission Control for Coexisting Heterogeneous Wireless Technologies”, IEEE Global Telecommunications Conference (GLOBECOM), New Orleans, USA, 2008.

[118] A. Rahmati and L. Zhong, “Context-Based Network Estimation for Energy-Efficient Ubiquitous Wireless Connectivity”, IEEE Transactions on Mobile Computing, vol. 10(1), pp. 54-66, 2011.

[119] A. Izquierdo and N. T. Golmie, “Improving Security Information Gathering with IEEE 802.21 to Optimize Handover Performance”, 12th international conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM), 2009, pp. 96-105.

[120] J. Zander, “Radio Resource Management in Future Wireless Networks: Requirements and Limitations”, IEEE Communications Magazine, vol. 35(8), pp. 30-36, 1997.

[121] D. Niyato and E. Hossain, “Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches”, IEEE Network, vol. 19(5), pp. 5-11, 2005.

[122] E. Patouni, N. Alonistioti and L. Merakos, “Modeling and Performance Evaluation of Reconfiguration Decision Making in Heterogeneous Radio Network Environments”, IEEE Transactions on Vehicular Technology, vol. 59(4), pp. 1887-1900, 2010.

[123] A. Galani, K. Tsagkaris and P. Demestichas, “Information Flow for Optimized Management of Spectrum and Radio Resources in Cognitive B3G Wireless Networks”, Springer Network and Systems Management Journal, vol. 18(2), pp. 125-149, 2010.

[124] D. N. Skoutas and A. N. Rouskas, “A Scheduling Algorithm with Dynamic Priority Assignment for WCDMA Systems”, IEEE Transactions on Mobile Computing, vol. 8(1), pp. 126-138, 2009.

[125] A. Saatsakis and P. Demestichas, “Context Matching for Realizing Cognitive Wireless Networks Segments”, Springer Wireless Personal Communications, vol. 55(3), pp. 407-440, 2010.

[126] M. D. Reuver and T. Haaker, “Designing Viable Business Models for Context-Aware Mobile Services”, Elsevier Telematics and Informatics, vol. 26(3), pp. 240-248, 2009.

[127] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies, “The Case for VM-based Cloudlets in Mobile Computing”, IEEE Pervasive Computing Journal, vol. 8(4), pp. 14-23, 2009.

[128] L. Badia, M. Levorato, F. Librino and M. Zorzi, “Cooperation Techniques for Wireless Systems from a Networking Perspective”, IEEE Wireless Communications, vol. 17(2), pp. 89-96, 2010.

[129] D. Riecken, “Personalized Views of Personalization”, Communications of the ACM, vol. 43(8), pp. 26–28, 2000.

[130] A. Jameson, “Modeling both the context and the user”, Personal and Ubiquitous Computing, vol. 5(1), pp. 29–33, 2001.

[131] G. Adomavicius and A. Tuzhilin, “Personalization Technologies: A Process-Oriented Perspective”, Communications of the ACM, vol. 48(10), pp. 83-90, 2005.

[132] R. L. Aguiar, A. Sarma, D. Bijwaard, L. Marchetti, and P. Pacyna, "Pervasiveness in a Competitive Multi-Operator Environment: The Daidalos Project", IEEE Communications Magazine, vol. 45(10), pp. 22-26, 2007.

[133] L. Hilty and M. Hercheui, "ICT and Sustainable Development," What Kind of Information Society? Governance, Virtuality, Surveillance, Sustainability, Resilience, Springer Boston, 2010, pp. 227-235.

[134] K. C. Laudon, "Ethical Concepts and Information Technology", Communications of the ACM, vol. 38(12), pp. 33-40, 1995.

[135] A. Radwan and J. Rodriguez, "Energy Saving in Multi-standard Mobile Terminals through Short-range Cooperation", EURASIP Journal on Wireless Communications and Networking, May 2012.

[136] Z. Hasan, H. Boostanimehr and V. K. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges", IEEE Communications Surveys & Tutorials, vol. 13(4), pp. 524-540, 2011.

[137] J. Saltzer and M. Schroeder, "The Protection of Information in Computer Systems", IEEE Computer Society Press, vol. 63(9), pp. 1278–1308, 1975.

[138] "Trust in the Information Society", A Report of the Advisory Board for Research & Innovation on Security, Privacy and Trustworthiness in the Information Society (RISEPTIS), 2010.

[139] A. P. Bianzino, C. Chaudet, D. Rossi and J. Rougier, "A Survey of Green Networking Research", IEEE Communications Surveys & Tutorials, vol. 14(1), pp. 3-20, 2012.

[140] Emin Islam Tatlı, "Security in Context-aware Mobile Business Applications", PhD Thesis, Mannheim University, 2009.

[141] Mark Stamp, "Information Security Principles and Practice", John Wiley & Sons Inc., ISBN: 978-0471738480, 2006.

[142] P. Abi-Char, A. M'hamed, B. EL-Hassan and M. Mokhtari, "Controlling Trust and Privacy in Context-Aware Environments, State of Art and Future Directions", IGI Publishing under Book Title: Trust Modeling and management in Digital Environments: From Social Concept to System Development, Book Edited by Nokia Research Center, pp. 352-377, Finland, 2010.

[143] E. Aivaloglou, S. Gritzalis and C. Skianis, "Requirements and Challenges in the Design of Privacy-aware Sensor Networks", 49th IEEE GLOBECOM Conference, San Francisco, USA, 2006.

[144] J. Cho, A. Swami, and I. Chen, "A Survey on Trust Management for Mobile Ad Hoc Networks", IEEE Communications Surveys & Tutorials, vol. 13(4), pp. 562-583, 2011.

[145] M. S. Lund, B. Solhaug and K. Stlen, "Evolution in Relation to Risk and Trust Management", IEEE Computer Society, vol. 43(5), pp. 49-55, 2010.

[146] O. E. Falowo and H. A. Chan, "Joint call admission control algorithms: Requirements, approaches and design considerations" Elsevier Computer Communications, vol. 31(6), pp. 1200-1217, 2008.

[147] Y. Fang and Y. Zhang, “Call admission control schemes and performance analysis in wireless mobile networks”, IEEE Transactions on Vehicular Technology, vol. 51(2), pp. 371-382, 2002.

[148] I. R. Chen, O. Yilmaz and I. L. Yen, “Admission control algorithms for revenue optimization with QoS guarantees in mobile wireless networks” Springer Wireless Personal Communications, vol. 38(3), pp. 357-376, 2006.

[149] B. Li, L. Li, K. M. Sivalingam and X-R Cao, “Call admission control for voice/data integrated cellular networks: performance analysis and comparative study” IEEE Journal on Selected Areas in Communications, vol. 22(4), pp. 706–718, 2004.

[150] S. E. Ogbonmwan and W. Li, “Multi-threshold bandwidth reservation scheme of an integrated voice/data wireless network”, Elsevier Journal on Computer Communications, vol. 29(9), pp. 1504-1515, 2006.

[151] Y. Fang, “Thinning Algorithms for Call Admission Control in Wireless Networks”, IEEE Transactions on Computers, vol. 52(5), pp. 685–687, 2003.

[152] J. Yao, J. W. Mark, T. C. Wong, Y. H. Chew, K. M. Lye and K-C. Chua, “Virtual partitioning resource allocation for multiclass traffic in cellular systems with QoS constraints”, IEEE Transactions on Vehicular Technology, vol. 53(3), pp. 847- 864, 2004.

[153] O. Yilmaz, I. R. Chen, G. Kulczycki and W. B. Frakes, “Performance Analysis of Spillover-Partitioning Call Admission Control in Mobile Wireless Networks. Springer Wireless Personal Communications, vol. 53(1), pp. 111-131, 2010.

[154] C. T. Chou and K. G. Shin, “Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service” IEEE Transactions on Mobile Computing, vol. 3(1), pp. 5–17, 2004.

[155] K. Zheng, F. Hu, W. Wang, W. Xiang and M. Dohler, “Radio resource allocation in LTE-advanced cellular networks with M2M communications” IEEE Communications Magazine, vol. 50(7), pp. 184-192, 2012.

[156] C. H. Yu, K. Doppler, C. B. Ribeiro and O. Tirkkonen, “Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks”, IEEE Transactions on Wireless Communications, vol. 10(8), pp. 2752-2763, 2011.

[157] T. Elkourdi and O. Simeone, “Femtocell as a Relay: An Outage Analysis”, IEEE Transactions on Wireless Communications, vol. 10(12), pp. 4204-4213, 2013.

[158] D. T. Hoang, D. Niyato and P. Wang, “Optimal admission control policy for mobile cloud computing hotspot with cloudlet”, IEEE Wireless Communications and Networking Conference (WCNC), 1-4/4, Paris, France, 2012.

[159] H. T. Dinh, C. Lee, D. Niyato and P. Wang, “A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches”, Wireless Communications and Mobile Computing (WCMC), doi: 10.1002/wcm.1203.

[160] L. Guan, X. Ke, M. Song and J. Song, “A Survey of Research on Mobile Cloud Computing”, IEEE/ACIS 10th International Conference on Computer and Information Science (ICIS), pp. 387-392, May 2011.

[161] P. Simoens, F. De Turck, B. Dhoedt and P. Demeester, “Remote Display Solutions for Mobile Cloud Computing”, IEEE Computer, vol. 44(8), pp. 46-53, 2011.

[162] B. G. Chun and P. Maniatis, “Dynamically Partitioning Applications Between Weak Devices and Clouds”, 1st ACM Workshop on Mobile Cloud Computing and Services (MCS), San Francisco, USA, 2010.

[163] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, “MAUI: Making Smartphones Last Longer with Code Offload”, ACM MobiSys 2010, Association for Computing Machinery, 2010.

[164] H. Liang, D. Huang, L. X. Cai, X. Shen and D. Peng, “Resource allocation for security services in mobile cloud computing”, Computer Communications Workshops in IEEE INFOCOM Conference, pp. 191-195, 2011.

[165] FP7 MCN Project, “Mobile Cloud Networking”, https://www.mobile-cloud-networking.eu/.

[166] FP7 iJoin Project, “Interworking and Joint Design of an Open Access and Backhaul Network Architecture for Small Cells based on Cloud Networks”, http://www.ict-ijoin.eu/.

[167] S-P. Yeh, S. Talwar, W. Geng, N. Himayat and K. Johnsson, “Capacity and Coverage Enhancement in Heterogeneous Networks”, IEEE Wireless Communications, vol. 18(3), pp. 32-38, 2011.

[168] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan and M. C. Reed, “Femtocells: Past, Present, and Future”, IEEE Journal on Selected Ares of Communications, vol. 30(3), pp. 497-508, 2012.

[169] Interference Management in OFDMA Femtocells, Small Cell Forum, March 2010.

[170] P. Xia, V. Chandrasekhar and J. Andrews, “Open vs. Closed Access Femtocells in the Uplink”, IEEE Transactions on Wireless Communications, vol. 9(12), pp. 3798-3809, 2010.

[171] W.C. Cheung, T.Q.S. Quek and M. Kountouris, “Throughput Optimization, Spectrum Allocation, and Access Control in Two-Tier Femtocell Networks”, IEEE Journal on Selected Areas of Communications, vol. 30(3), pp. 561-574, 2012.

[172] I. F. Akyildiz, D. M. Gutierez-Estevez and E. C. Reyes, “Femto-Relay Systems and Methods of Managing Same”, US patent US 2012/0076027 A1, March 2012.

[173] G. de la Roche, A. Valcarce, D. López-Pérez and J. Zhang, “Access Control Mechanisms for Femtocells”, IEEE Communications Magazine, vol. 48(1), pp. 33-39, 2010.

[174] C. Hoymann, W. Chen, J. Montojo, A Golitschek, C. Koutsimanis and X. Shen, “Relaying Operation in 3GPP LTE: Challenges and Solutions”, IEEE Communications Magazine, vol. 50(2), pp. 156-162, 2012.

[175] N. Nomikos, P. Makris, D. N. Skoutas, D. Vouyioukas and C. Skianis, “A Cooperation Framework for LTE Femtocells’ Efficient Integration in Cellular Infrastructures Based on Femto Relay Concept”, IEEE International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD), Barcelona, Spain, September 2012.

[176] A. Bletsas, H. Shin, and M. Z. Win, “Cooperative Communications with Outage-Optimal Opportunistic Relaying”, IEEE Transactions on Wireless Communications, vol. 6(9), pp. 3450-3460, 2007.

[177] A. Rath, S. Hua and S. S. Panwar, “FemtoHaul: Using Femtocells with Relays to Increase Macrocell Backhaul Bandwidth”, INFOCOM IEEE Conference on Computer Communications Workshops, pp. 1-5, March 2010.

[178] Z. Dizhi and S. Wei, “Interference-Controlled Load Sharing with Femtocell Relay for Macrocells in Cellular Networks”, Global Telecommunications Conference (GLOBECOM), pp. 1-5, Dec. 2011.

[179] A. Tyrrell, F. Zdarsky, E. Mino and M. Lopez, “Use Cases, Enablers and Requirements for Evolved Femtocells”, Vehicular Technology Conference (VTC Spring), pp. 1-5, May 2011.

[180] A. Adhikary, V. Ntranos and G. Caire, “Cognitive femtocells: Breaking the spatial reuse barrier of cellular systems”, Information Theory and Applications Workshop (ITA), pp. 1-10, Feb. 2011.

[181] 3GPP, Technical Specifications Group Services and Systems Aspects; End-to-end Quality of Service (QoS) Concept and Architecture. 3GPP TS 23.107, V9.1.0, Release 9, 2010.

[182] 3GPP, Technical Specifications Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects. 3GPP TR 36.814, V9.0.0, Release 9, 2010.

[183] S. Sesia, I. Toufik and M. Baker, “LTE The UMTS Long Term Evolution, From Theory to Practice”, Wiley, New York, 2009, ISBN:978-0-470-69716-0.

[184] J. N. Laneman, D.N.C. Tse and G. W. Wornell, “Cooperative diversity in wireless networks: efficient protocols and outage behavior”, IEEE Transactions on Information Theory, vol. 50(12), pp. 3062–3080, 2004.

[185] H. Holma and A. Toskala, “LTE for UMTS: OFDMA and SC-FDMA Based Radio Access”, John Wiley & Sons, 2009.

[186] P. Stuckmann and R. Zimmermann, “European research on future Internet design”, IEEE Wireless Communications, vol. 16(5), pp. 14-22, 2009.

[187] J. Schonwalder, M. Fouquet, G. Rodosek and I. Hochstatter, “Future Internet = content + services + management”, IEEE Communications Magazine, vol. 47(7), pp. 27-33, 2009.

[188] W. Song, W. Zhuang and Y. Cheng, “Load balancing for cellular/WLAN integrated networks”, IEEE Network magazine, vol. 21(1), pp. 27-33, 2007.

[189] M. Dohler, T. Watteyne, and J. Alonso, “Machine-to- Machine: An Emerging Communication Paradigm”, Tutorial, PIMRC 2010, 26 Sept. 2010, Istanbul, Turkey; also at Globecom 2010, Miami 2010, USA.

[190] K. Zheng, F. Hu, W. Wang, W. Xiang and M. Dohler, “Radio resource allocation in LTE-advanced cellular networks with M2M communications”, IEEE Communications Magazine, vol. 50(7), pp. 184-192, 2012.

[191] F. Zarai, K. B. Ali, M. S. Obaidat and L. Kamoun, “Adaptive call admission control in 3GPP LTE networks”, International Journal of Communication Systems, doi: 10.1002/dac.2415, 2012.

[192] T. Taleb and A. Ksentini, “QoS/QoE Predictions-based Admission Control for Femto Communications”, IEEE International Conference on Communications (ICC), Ottawa, Canada, 2012.

[193] C. Olariu, J. Fitzpatrick, P. Perry and L. Murphy, "A QoS based call admission control and resource allocation mechanism for LTE femtocell deployment", IEEE Consumer Communications and Networking Conference (CCNC), pp. 884-888, 2012.

[194] O. Tipmongkolsilp, S. Zaghloul and A. Jukan, "The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends", IEEE Communications Surveys & Tutorials, vol. 13(1), pp. 97-113, 2011.

[195] M. Z. Chowdhurry et al., "Dynamic SLA Negotiation using Bandwidth Broker for Femtocell Networks", 1st International Conference on Ubiquitous and Future Networks (ICUFN), pp. 12-15, 2009.

[196] 3GPP, LTE, "Service Requirements for Home NodeBs (UMTS) and eNodeBs (LTE)", TS 22.220, V.10.3.0, Release 10, 2010.

[197] S. Saunders, S. Carlaw, A. Giustina, R. R. Bhat, V. S. Rao and R. Siegberg, "Femtocells: Opportunities and Challenges for Business and Technology", John Wiley & Sons Ltd, 2009.

[198] Small Cell forum White Paper, "Integrated Femto-WiFi (IFW) Networks", published 28 Feb 2012, accessible at: smallcellforum.org/smallcellforum_resources/pdfsend01.php?file=Integrated-Femto-WiFi-Networks-White-Paper_sm.pdf

[199] K. Samdanis, T. Taleb and S. Schmid, "Traffic Offload Enhancements for eUTRAN", IEEE Communications Surveys & Tutorials, vol. 14(3), pp. 884-896, 2012.

[200] S. F. Hasan, N. H. Siddique and S. Chakraborty, "Femtocell versus WiFi – A Survey and Comparison of Architecture and Performance", 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), pp. 916-920, 2009.

[201] C. H. Ko and H. Y. Wei, "On-Demand Resource Sharing Mechanism Design in Two-Tier OFDMA Femtocell Networks", IEEE Transactions on Vehicular Technology, vol. 60(3), pp. 1059-1071, 2011.

[202] D. Knisely, T. Yoshizawa and F. Favichia, "Standardization of Femtocells in 3GPP", IEEE Communications Magazine, vol. 47(9), pp. 68-75, 2009.

[203] A. Rouskas, D. N. Skoutas, G. Kormentzas and D. D. Vergados, "Code Reservation Schemes at the Forward Link in WCDMA" Elsevier Computer Communications, vol. 27(9), pp. 792-800, 2004.

[204] R. Skehill, M. Barry, W. Kent, M. O'Callaghan, N. Gawley and S. Mcgrath, "The Common RRM Approach to Admission Control for Converged Heterogeneous Wireless Networks" IEEE Wireless Communications, vol. 14(2), pp. 48-56, 2007.

[205] E. Stevens-Navarro, A. H. Mohsenian-Rad and V. Wong, "Connection Admission Control for Multiservice Integrated Cellular/WLAN System" IEEE Transactions on Vehicular Technologies, vol. 57(6), pp. 3789-3800, 2008.

[206] D. Calabruig, J. F. Monserrat, D. M. Sacristan and N. Cardona, "Joint dynamic resource allocation for QoS provisioning in multi-access and multi-service wireless systems", Springer MONET, vol. 15(5), pp. 627-638, 2010.

[207] P. Kulkarni, W. H. Chin and T. Farnham, "Radio Resource Management Considerations for LTE Femto Cells", ACM SIGCOMM Computer Communication Review, vol. 40(1), pp. 26-30, 2010.

[208] http://www.cisco.com/en/US/prod/collateral/wireless/ps11035/ps11047/ps11072/ps12542/data_sheet_c78-712213.html, accessed June 2013.

[209] http://www.netgear.com/service-provider/products/mobile-broadband/3G-smallcell/DEVG2000F.aspx#, accessed June 2013.

[210] http://www.alcatel-lucent.com/products/9360-small-cell, accessed June 2013.

[211] http://www.nec.com/en/global/solutions/nsp/femto_sc/enterprise.html, accessed June 2013.

[212] http://www.astri.org/main/?contentnamespace=technologies:ct:communications_software:lte_femtocell, accessed June 2013.

[213] http://www.btiwireless.com/products/small-cells/, accessed June 2013.

[214] http://www.lever.co.uk/ruckus-smartcell-gateway-hetnet-cell-3g-offload-4g-3gpp-wag-i-wlan.htm, accessed June 2013.

[215] EXALTED Project Deliverable 2.3, "The EXALTED System Architecture", Available online: www.ict-exalted.eu, Aug. 2012.

[216] V. Misic, J. Misic, X. Lin and D. Neradzic, "Capillary Machine-to-Machine Communications: The Road Ahead", X. Y. Li, S. Papavasiliou, S. Ruehrup (Eds.): ADHOC-NOW, Springer LNCS 7363, pp. 413-423, 2012.

[217] G. Wu, S. Talwar, K. Johnsson, N. Himayat and K. D. Johnson, "M2M: From mobile to embedded internet," IEEE Communications Magazine, vol. 49(4), pp. 36-43, 2011.

[218] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao and M. Guizani, "Home M2M networks: Architectures, standards, and QoS improvement," IEEE Communications Magazine, vol. 49(4), pp. 44-52, 2011.

[219] X. G. Wang, G. Min, J. E. Mellor, K. Al-Begain and L. Guan, "An adaptive QoS framework for integrated cellular and WLAN network", Elsevier Computer Networks, vol. 47(2), pp. 167-183, 2005.

[220] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture for differentiated services", Internet RFC 2475, 1998.

[221] B. Statovci-Halimi and G. Franzl, "QoS Differentiation and Internet Neutrality - A Controversial Issue within the Future Internet Challenge", Springer Telecommunication Systems, 2011, doi:10.1007/s11235-011-9517-1.

[222] 3GPP TS 22.368 v12.1.0, "Service requirements for Machine-Type Communications (MTC) Stage 1", Release 12, Dec. 2012.

[223] ETSI TS 102 689, v1.1.1, "Machine-to-Machine Communications (M2M): M2M Service Requirements," Aug. 2010.

[224] V. B. Misic, J. Misic and D. Nerandzic, "Extending LTE to support machine-type communications", IEEE International Conference on Communications (ICC), pp. 6977-6981, June 2012.

[225] 3GPP TR 23.888 v11.0.0, “System Improvements for Machine-Type Communications (MTC)”, Release 11, Sept 2012.

[226] T. Taleb and A. Kunz, “Machine type communications in 3GPP networks: potential, challenges, and solutions”, IEEE Communications Magazine, vol. 50(3), pp. 178-184, 2012.

[227] A. Amokrane, A. Ksentini, Y. Hadjadj-Aoul and T. Taleb, “Congestion control for machine type communications”, IEEE International Conference on Communications (ICC), pp. 778-782, June 2012.

[228] A. G. Gotsis, A. S. Lioumpas and A. Alexiou, “M2M Scheduling over LTE: Challenges and New Perspectives”, IEEE Vehicular Technology Magazine, vol. 7(3), pp. 34-39, 2012.

[229] A. Gotsis, A. Lioumpas and A. Alexiou, “Evolution of Packet Scheduling for Machine-Type Communications over LTE: Algorithmic Design and Performance Analysis”, IEEE Globecom’12, Anaheim, CA, USA, Dec. 2012.

[230] S. Y. Lien, K. C. Chen and Y. Lin, “Toward ubiquitous massive accesses in 3GPP machine-to-machine communications”, IEEE Communications Magazine, vol. 49(4), pp. 66-74, 2011.

[231] R. Liu, W. Wu, H. Zhu and D. Yang, “M2M-Oriented QoS Categorization in Cellular Network”, 7th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Sept. 2011.

[232] M. Jonckheere and J. Mairesse, “Towards an erlang formula for multiclass networks”, Queueing Systems Theory Appl., vol. 66(1), pp. 53-78, 2010.

[233] M. Armbrust et al., “Above the clouds: A Berkeley View of Cloud Computing”, February 2009.

[234] W. Jansen and T. Grance, “Guidelines on Security and Privacy in Public Cloud Computing”, NIST: National Institute of Standards and Technology, Technical Report 800-144, 2011.

[235] S. Sakr, A. Liu, D. Batista and M. Alomari, “A Survey of Large Scale Data Management Approaches in Cloud Environments”, IEEE Communications Surveys & Tutorials, vol. 13(3), pp. 311-336, 2011.

[236] M. R. Head, A. Kochut, C. Schulz and H. Shaikh, “Virtual Hypervisor: Enabling Fair and Economical Resource Partitioning in Cloud Environments”, IEEE Network Operations and Management Symposium (NOMS), pp.104-111, 2010.

[237] J. C. Mace, A. V. Moorsel, and P. Watson, “The case for dynamic security solutions in public cloud workflow deployments”, IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W), pp. 111-116, June 2011.

[238] A. Kochut and K. A. Beaty, “On Strategies for Dynamic Resource Management in Virtualized Server Environments”, 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), pp. 193-200, 2007.

[239] A. Waller, I. Sandy, E. Power, E. Aivaloglou, C. Skianis, A. Muñoz and A. Mana, “Policy Based Management for Security in Cloud Computing”, 1st International Workshop on Security & Trust for Applications in Virtualised Environments, J. Lopez, (Ed.), June 2011, Loutraki, Greece, Springer CCIS.

[240] P. T. Endo et al., "Resource Allocation for Distributed Cloud: Concepts and Research Challenges", IEEE Network, vol. 25(4), pp. 42-46, July-August 2011.

[241] H. H. Chen and M. Guizani, "Next Generation Wireless Systems and Networks", John Wiley & Sons Ltd., 2006.

[242] B. Saovapakhiran and M. Devetsikiotis, "Enhancing Computing Power by Exploiting Underutilized Resources in the Community Cloud", IEEE International Conference on Communications (ICC), June 2011.

[243] FP7 ICT PASSIVE Project, Policy-Assessed system-level Security of Sensitive Information processing in Virtualised Environments, http://ict-passive.eu/.

[244] K. Kumar and Y. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?", IEEE Computer, vol. 43(4), pp. 51-56, 2010.

[245] R. Urgaonkar, U. C. Kozat, K. Igarashi and M. J. Neely, "Dynamic Resource Allocation and Power Management in Virtualized Data Centers", IEEE Network Operations and Management Symposium (NOMS), pp. 479-486, April 2010.

[246] B. Urgaonkar and P. Shenoy, "Sharc: Managing CPU and Network Bandwidth in Shared Clusters", IEEE Transactions on Parallel and Distributed Systems, vol. 15(1), pp. 2–17, 2004.

[247] M. Steinder, I. Whalley and D. Chess, "Server Virtualization in Autonomic Management of Heterogeneous Workloads", ACM SIGOPS Operating Systems Review, vol. 42(1), pp. 94-95, 2008.

[248] X. Wang, Z. Du, Y. Chen and S. Li, "Virtualization-Based Autonomic Resource Management for Multi-Tier Web Applications in Shared Data Center", Journal of Systems and Software, vol. 81(9), pp. 1591–1608, 2008.

[249] B. Rochwerger et al., "Reservoir - When One Cloud Is Not Enough", Computer, vol. 44(3), pp. 44-51, March 2011.

[250] R. Craig et al., "Cloud Computing in the Public Sector: Public Manager's Guide to Evaluating and Adopting Cloud Computing", CISCO Internet Business Solutions Group, November 2009, http://www.cisco.com/web/about/ac79/docs/wp/ps/Cloud_Computing_112309_FINAL.pdf

[251] D. Wan, A. Greenway, J. G. Harris and A. E. Alter, "Six Questions Every Health Industry Executive Should Ask About Cloud Computing", Accenture Institute for Health & Public Service Value, 2010, http://newsroom.accenture.com/images/20020/HealthcareCloud.pdf

[252] A. Beloglazov and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Center", 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGRID), pp. 577-578, 2010.

[253] D. Ardagna, B. Panicucci, M. Trubian and L. Zhang, "Energy-Aware Autonomic Resource Allocation in Multi-Tier Virtualized Environments", IEEE Transactions on Services Computing, vol. 5(1), pp. 2-19, 2012.

[254] M. Sedaghat, F. Hernandez and E. Elmroth, "Unifying Cloud Management: Towards Overall Governance of Business Level Objectives", 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 591-597, May 2011.

[255] I. Goiri, J. Guitart and J. Torres, “Characterizing Cloud Federation for Enhancing Providers' Profit”, 3rd IEEE International Conference on Cloud Computing (CLOUD), pp.123-130, July 2010.

[256] R. Buyya, R. Ranjan and R. N. Calheiros, “InterCloud: Utility-Oriented Federation of Cloud Computing Environments for Scaling of Application Services”, 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Springer LNCS 6081, pp. 13-31, 2010.

[257] M. Jonckheere and J. Mairesse “Towards an Erlang formula for multiclass networks”, Queueing Systems Theory Applications, vol. 66(1), pp. 53-78, 2010.

[258] A. J. Ferrer et al., “OPTIMIS: a Holistic Approach to Cloud Service Provisioning”, Future Generation Computer Systems Journal, vol. 28(1), pp. 66-77, 2012.