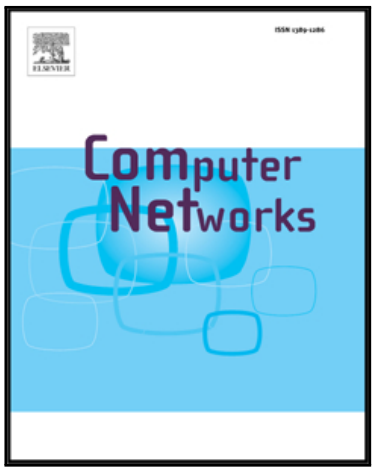

Context-aware, user-driven, network-controlled RAT selection for 5G networks

Sokratis Barmpounakis , Alexandros Kaloxylos ,
Panagiotis Spapis , Nancy Alonistioti

PII: S1389-1286(16)30423-6
DOI: 10.1016/j.comnet.2016.12.008
Reference: COMPNW 6071

To appear in: *Computer Networks*

Received date: 28 July 2016
Revised date: 12 December 2016
Accepted date: 15 December 2016

Please cite this article as: Sokratis Barmpounakis , Alexandros Kaloxylos , Panagiotis Spapis , Nancy Alonistioti , Context-aware, user-driven, network-controlled RAT selection for 5G networks, *Computer Networks* (2016), doi: 10.1016/j.comnet.2016.12.008

Context-aware, user-driven, network-controlled RAT selection for 5G networks

Sokratis Barmpounakis[a,*] sokbar@di.uoa.gr, Alexandros Kaloxylos[b], Panagiotis Spapis[b], Nancy Alonistioti[a]

[a]University of Athens, Athens, Greece

[b]Huawei ERC, Munich, Germany

[*]Corresponding author.

***Abstract*— It is expected that in the very near future, cellular networks will have to deal with a massive data traffic increase, as well as a vast number of devices. Optimal placement of the end devices to the most suitable access network is expected to provide the best Quality of Service (QoS) experience to the users but also the maximum utilization of the scarce wireless resources by the operators. Several on-going proposals attempt to overcome the existing barriers by enabling the use of Wi-Fis and femto-cells to cater for part of the load generated by the end devices. The evolution of the Access Network Discovery and Selection Function (ANDSF) for the core part of the cellular network, as well as the Hotspot 2.0 approach, are currently being subject to thorough discussions and studies and are expected to facilitate a seamless 3GPP-WiFi interworking. During the past years, several Radio Access Technology (RAT) selection schemes have been proposed. However, these schemes do not take into consideration the opportunities offered by these new standardized approaches. Our paper acts in a manifold way: Firstly, it proposes COmpAsS, a Context-Aware RAT Selection mechanism, the main part of which operates on the User Equipment (UE)-side, minimizing signaling overhead over the air interface and computation load on the base stations. Secondly, we discuss in detail the architectural perspective; i.e., the extensions needed in the network interfaces that enable the exchange of the required context information among the respective network entities and in accordance with the 3GPP trends in relation to the context-aggregating entities. Furthermore, we quantify the signaling overhead of the proposed mechanism by linking it to the current 3GPP specifications and performing a comprehensive per-parameter analysis. Finally, we evaluate the novel scheme via extensive simulations in a complex and realistic 5G use case, illustrating the clear advantages of our approach in terms of key QoS metrics, i.e. the user-experienced throughput and delay, both in the uplink and the downlink.**

Keywords—5G network, access network selection, ANDSF, Hotspot 2.0

I. INTRODUCTION

Effective and efficient network planning is essential to deal with the constantly increasing number of mobile users and bandwidth-intensive services. Already, operators attempt to address this challenge in a manifold manner, i.e., by increasing capacity with new radio spectrum [1] adding multi-antenna techniques [2], implementing more efficient modulation and coding schemes [3],[4], etc. However, even these solutions alone are insufficient in the most crowded environments, as well as at cell edges where performance can significantly degrade. As a result, one of the main directions that operators are also heading towards is "densifying" the existing network environments, leading to the Ultra Dense Networks (UDNs). According to network traffic data analysis and projections [5], one of the greatest challenges in the forthcoming wireless networks era is that 5G networks will have to cope with a huge increase both in terms of data traffic as well as number of end devices (e.g., smartphones, tablets, sensors etc.). Besides the tremendous growth, which is expected in terms of number of devices, due to an increasingly diverse set of new and yet unforeseen services, users and applications (including machine-to-machine modules, smart cities, industrial automation, etc.), novel and less predictable mobile traffic patters are also expected to emerge [6]. In order to address this issue the research community designs solutions to improve the spectral efficiency, to increase the network cell density and to exploit the underutilized radio spectrum resources [7]. One of the main trends suggests the exploitation of the available small cells, i.e. primarily femto-cells (Home eNBs) or Wi-Fi Access Points (APs) to efficiently distribute the network load [8][9] via intelligent dynamic steering of the network traffic.

It is envisaged that the aforementioned trend will inevitably result in very dense deployments, in which on the one hand, Long Term Evolution (LTE) base stations (BSs) will co-exist with their 5th generation evolution, while in addition, 3GPP networks will co-exist with the non-3GPP ones (primarily Wi-Fi), creating thus a multi-tier architecture consisting of heterogeneous radio access technologies. Some of the greatest challenges in such dense wireless environments are the efficient inter-working between the legacy with the latest cellular systems, as well as with Wi-Fi APs, the optimization of the UE placement - RAT selection procedures, as well as the minimization of the unnecessary handovers -and ping-pong effect-related events- between adjacent RATs and cells, which inevitably deteriorate the provided QoS to the users: The handover procedure in the current Evolved Packet Core (EPC)/4G system involves latency overheads, even in limited coverage areas. over the GPRS Tunnelling Protocol (GTP) tunnel [10]. In order to enable seamless UE mobility when moving across the different (H)eNBs, the S-GW (at the network core) communicates with the eNBs (at the network edge) to perform handover management; QoS allocation, traffic condition monitoring, user terminal mobility management and security tasks are also forwarded to the Packet Gateway (P-GW). At the same time, the eNB, the S-GW, and the P-GW perform several signaling procedures to handle the session setup at different levels. Such an approach decreases considerably the network performance by increasing the latency and thereby reducing the QoS required for the future real-time applications. Thus, it becomes of utmost importance that frequent or unnecessary handovers in such ultra-dense network environments are minimized; latency overheads should be minimized and the optimal RAT options for the UEs should be available in an efficient way via a viable RAT selection approach.

Lately, new directions have been presented by 3GPP's specification groups [11] towards the network capacity issue optimization and the so called tight interworking of 3GPP and non-3GPP access technologies, with plenty of these novel directions and standards planned to be integrated in the forthcoming releases. In relation to the efficient interworking between heterogeneous wireless systems (e.g., LTE and Wi-Fi), although during the last decade there has been considerable progress in terms of specifications and standards, still a successful demonstration of a seamless integration of Wi-Fi APs with commercial cellular networks, and in realistic 5G business cases is missing [12]-[14]. This is because of a number of reasons. Wi-Fi suffers from interference issues since it operates on the unlicensed spectrum. Most importantly however, switching from a cellular network to a Wi-Fi access point has not properly yet evolved to a fully transparent process -from different perspectives-; for the end users the authentication process had to take place manually –thus, deteriorating the QoE-; furthermore, the mobility of multiple flows (even of the same service) among different PDN connections and access technologies was only recently standardized and described [15]. In addition to the first point –and as this is described by Hotspot 2.0-, all the "islands" of hotspots should be also interconnected into larger "footprints" via further roaming agreements between Wi-Fi operators. Finally, there are still diverse strategies in the way non-3GPP networks are handled by different devices and operating systems, meaning that the software that handles the active UE connections (e.g. the "Connectivity Manager" in Android) has not been standardized. In some cases, even the same OS handles differently the connections depending on the version of the OS, e.g. [16].

The co-existence of LTE and Wi-Fi has also been studied in the literature; the key inferences identified in such efforts [17]-[18] also confirm the aforediscussed potentials along with the issues and limitations that accompany the tighter coupling. In [17] the authors describe the numerous degrees, in which the Radio Access Network (RAN)-level integration may improve the LTE/Wi-Fi cooperation in the context of the emerging Heterogeneous Networks (HetNets). At the same time, however, they highlight the fact that novel multi-RAT and multi-tier solutions require additional infrastructure enablers, such as network management interfaces, able to deliver flexible core network connectivity for the envisioned system architecture of next-generation 5G systems. According to the authors in [18], the LTE-WiFi interworking is a promising solution, however, due to its considerable drawbacks (i.e., signal attenuation though walls, CSMA/CA limitations allowing only one link to be active at one time, etc.), supplementary solutions could be considered such as optical wireless communication systems, such as LiFi [19]. Still, they also conclude that big standardization efforts are required in order to define new modes of simultaneous transmissions to multiple users, adding also the challenge of the complexity limits with larger number of antennas.

Optimizing the traffic steering strategy among dense heterogeneous network scenarios is one of the compelling challenges in the latest releases to be addressed, as well. Although, already before Rel.12, the Wi-Fi interworking was supported at the core-network level (via IP-session continuity supported by GTP, MIP and PMIP protocols [10],[20]), the Wi-Fi and Public Land Mobile Network (PLMN) selection mechanisms were not satisfying enough to be

deployed, and the traffic steering mechanisms via ANDSF [22] were not sufficient. What was missing was a mechanism to determine the access network to use for a given IP flow. In the current release, ANDSF has been enhanced introducing the Inter-System and Inter-APN Routing Policy (ISRP and IARP) rules [21],[22], which enable the UE to determine the access network to route its active flows, by taking into account besides the Received Signal Strength Indicator (RSSI) levels, the Wi-Fi BS load as well. Our mechanism that will be thoroughly described in the following sections, not only does utilize the ANDSF; it takes the aforementioned one step further, creating an even more holistic picture of the network conditions of the candidate RATs in a lightweight and efficient manner.

In parallel with the steps being made in terms of the heterogeneous networks tight integration and interworking, RAT selection optimization for active UE sessions, i.e. handover procedure, is of paramount importance. Due to the fact that the selection of the access technology influences on a great extent both the resource allocation per user, as well as the generated interference among cells, the experienced quality of the respective services may vary greatly. Numerous schemes have been proposed to optimize the RAT selection procedure. These schemes extend from very simple solutions that often do not attempt to acquire a holistic picture of the network environment context (in order to avoid signaling overhead issues) to complex frameworks, which however require major modifications in the core network components, proposing diverse parameters as inputs towards the assessment of the available RAT choices. From the existing solutions' perspective, this leads to a considerable gap, which should accommodate solutions that will manage to utilize on the one hand all the existing information, –which is already being exchanged among the EPC network entities-, as well as take advantage of the forthcoming 3GPP releases' new features-, in order to optimize the tradeoff between context acquisition and signaling overhead and provide a feasible and realistic approach for efficient RAT selection in 5G. In addition to the previous, it must be also highlighted that with latency being one of the most crucial QoS metrics for next generation networks, any mechanism that will be finally deployed –either on the network or the UE side- should demonstrate exceptional performance in terms of algorithm execution/computation delays.

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

(a) This work concentrates on the context acquisition process: A comprehensive analysis on the network sources, respective interfaces and context information item types is made. In addition, an analytical approach is presented, which provides detailed insights on the information items, which are used, along with signaling overhead required to aggregate them. To the best of our knowledge, there is no previous work, which attempts to quantify the signaling overhead of the proposed context-based mechanism, as we will present in subsection IIIE.

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

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

A proof of concept version of a "lite" and simpler scheme has been validated in two earlier works [23],[24]. During the early evaluation of (parts of) this scheme in the aforementioned works, we evaluated an initial, basic rules set, upon which the fuzzy logic scheme was based, while we also performed some very basic simulations to initially assess the validity of the proposal. In this work, the algorithm of COmpAsS has been carefully enriched from numerous perspectives, which are explained in detail: the signaling cost perspective of the mechanism has been thoroughly investigated; there is always a crucial trade-off for context-based scheme, and this is the first time an assessment is being made relating taking also in account the current (3GPP's) available information items as well; from the algorithmic perspective, we have optimized the mechanism of the triggering events: an optimized *Threshold* and *Hysteresis* mechanism (see section III), that optimize its functionality from the energy consumption, as well as network signaling perspective; furthermore, the rules set that we apply on the fuzzy inference system has been progressively fine-tuned and optimized after numerous simulations and feedback loops assessing pre-defined Key

Performance Indicators (KPIs); next, the environment of the simulation has become more realistic and sophisticated, having also added a building propagation model, as well as a shadowing loss model, -which were previously missing-; last but not least, in this comprehensive round of simulations, we assign multiple traffic types to the UEs (VoIP, FTP, etc.), studying the behavior of the scheme and the fine-tuned rules set for completely diverse different types of IP flows (and QoS requirements respectively). In the previous -proof of concept- works, all the assumptions were made (regarding the required context information) without attempting to take into account the overhead imposed by the context acquisition process; the mechanism lacked several key configurable parameters; the rules of the fuzzy inference module were initially set without having been fine-tuned based on actual results and dedicated testing; last but not least, the simulation environment was much simpler (buildings-walls attenuation, signal propagation models, UE mobility models, only one traffic flow type, etc.).

The rest of the paper is organized as follows. In Section 2, which is divided in two main sub-sections, we initially present the standardization efforts in relation to RAT selection and Wi-Fi integration in cellular, while in the second sub-section we present the related work from the literature that attempts to deal with the aforementioned challenges. Section 3 is split into four main parts: the first presents an overview of the proposed scheme; the second sub-section goes through a comprehensive description of the mechanism; the next part is related to the fuzzy logic component of the system, the inputs and the outputs of the (de)fuzzification processes, as well as a number of surface figures that illustrate some use case examples of varying inputs; the final part of Section 3 focuses on the architectural perspective of the solution, the required interface extensions, as well as the functional requirements of the involved network entities. In Section 4, extensive simulation results based on a realistic 5G business case are presented. In Section 5, we summarize and draw the conclusions that are derived from the overall work presented and we discuss our future steps.

II. Related Work

In this section, we present the existing RAT selection policies and mechanisms specified in the 3GPP standards as well as in current literature. We identify the shortcomings and weaknesses of each contribution and conclude the section by introducing our proposed mechanism in Section 3.

A. Standardization efforts

The integration of Wi-Fi APs with the EPC has already been specified by 3GPP [12]-[14]. Novel technological solutions such as Hotspot 2.0 initiative from Wi-Fi Alliance and the Access Network Discovery and Selection function (ANDSF) -if combined together- may prove the right solution for simplifying the access of end users among RATs [25]. Furthermore, S2-a based mobility over GPRS Tunneling Protocol (SaMOG) [14] integrates trusted Wi-Fi APs to the EPC network. Additionally, roaming among cellular operators and wireless Internet service providers may also be supported. The new business case for cellular operators would be to maintain the same QoE for their services among different RATs, even if these RATs belong to different operators. Thus, the integration of cellular networks with Wi-Fi APs needs to be revised: both due to the new business cases that arise, as well as because of the novel, RAT selection-related protocols.

Hotspot 2.0 standard from Wi-Fi Alliance improves the ability of WLAN devices to discover and connect in a secure way to public Wi-Fi APs. Hotspot 2.0 builds on 802.11u specifications [26] that enable devices to discover information about the available roaming partners using query mechanisms. The Access Network Query Protocol (ANQP) [26] is the query and response protocol, which supports Hotspot 2.0. ANQP, apart from the operator's domain name, the accessible roaming partners, the type of the access point (private, public free, public chargeable, etc.), is capable of collecting the load information (i.e., total number of currently associated devices to the AP, channel utilization percentage and an estimate of the remaining available admission capacity). With load being one of the most crucial parameters linked with the QoS provided by a BS or Wi-Fi AP, this feature is of upmost importance when it comes to the evaluation and selection of the most appropriate RAT.

SaMOG [14] allows UEs to seamlessly handover between cellular and Wi-Fi network. According to SaMOG specification, the Wi-Fi gateway does not connect directly to the EPC via the Packet Gateway (PGW). Another network entity, the Trusted Wireless Access Gateway (TWAG) is used, acting as the perimeter security entity of the EPC network and connects to the PGW over a secure GTP tunnel.

ANDSF [22] is a cellular technology standard -closely coupled with the Policy and Charging Rules Function (PCRF) [27]- that implements dynamic data offload for the User Equipment (UE) in a structured method, while in addition,

enables the operator to store its policies for discovery and selection of RATs on a server. The UEs are updated with these polices by the server. The policies within ANDSF contain information on which of the available Wi-Fi hotspots are preferable during a specific time or day, and at a specific location as well, based on indications from past measurements.

The ANDSF Management Object (MO) is the primary representation of ANDSF. ANDSF MO may contain information with regard to the UE location, Inter-System Mobility Policies (ISMPs) and Inter-System Routing Policies (ISRPs) [28]. The ISRPs are available for UEs, which support IP Flow Mobility (IFOM), multiple-access Packet Data Network (PDN) connectivity (MAPCON), or non-seamless offload [29]-[31]. MAPCON enabled UEs may establish different PDN connections through different RATs. IFOM enabled terminals may establish a single PDN connection via multiple access networks, for instance 3G/LTE and Wireless Local Area Network (WLAN). For such UEs, IFOM enables to move individual IP flows from one access network to another with session continuity. The ANDSF prioritized rules in the case of MAPCON apply per PDN connections, while in IFOM and non-seamless offload cases per flow. ANDSF communicates with the UE over the S14 reference point.

The WLAN_NS working item of 3GPP [33] is working to Enhance 3GPP solutions for WLAN and access network selection based on Hotspot 2.0 and ensure that data, i.e., Management Objects (MO) and policies provided via HotSpot 2.0 and ANDSF are consistent. This alignment of ANDSF and HotSpot 2.0 provides an excellent basis for the complementarity of ANDSF and Hotspot 2.0, as well a number of multi-operator scenarios that can be supported. In [8], a rather exhaustive list of possible scenarios is presented.

From the analysis above it is clear that current standards and solutions pave the way to the operators to deploy flexible solutions where data flow can be handled, even on a per session basis, using various RATs. Also, much effort has been spent to communicate an operator's RAT selection policies to UEs. Although we have highlighted many key enablers from the standardization efforts that have evolved recently, there are still open problems that the industry needs to address before Wi-Fi/Cellular integration can be fully realized. These technical enablers on the one hand facilitate the design and development of a new generation of RAT selection mechanisms, -something required for the realization of 5G networks and their high QoS requirements-, however, they need to be federated by novel mechanisms that will bridge the gap between these integration efforts and the intelligent handling of the 5G HetNets. In particular, many of the challenges facing Wi-Fi/Cellular integration have to do with realizing a complete intelligent network selection solution that allows operators to steer traffic in a manner that maximizes user experience and addresses some of the challenges at the boundaries between RATs.

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

In this paper we take advantage of these enablers in order to –on the one hand- design the novel interfaces that are needed between the network entities in order to obtain minimal but valuable context information and –on the other- evaluate a real-time context-aware mechanism that operates in heterogeneous wireless network environments optimizing the RAT selection process in a dynamic manner, based on real-time network information and other critical context information.

B. Literature proposals

There has been a lot of effort into further optimizing the standardized mechanisms, and plenty of proposals and algorithmic solutions to improve the handover procedure.

The survey in [34] provides an overview of the main handover (HO) decision criteria in the current literature and presents a classification of existing HO decision algorithms for femto-cells. According to this [34], some researchers focus on evaluating the Reference Signal Received Power and/or Reference Signal Received Quality (RSRP/RSRQ) [35], the user location or speed, the mobility patterns, the battery level, the mean UE transmit power and the UE power consumption, the load of the cell and the service type. Apart from the case of RSRP, typically researchers are using multiple criteria (e.g., battery lifetime, traffic type, cell load, speed) and are using different tools (e.g., cost based functions, fuzzy logic, etc.) to reach a decision.

Xenakis et al. [36] presents the state of the art on vertical handover (VHO) mechanisms highlighting the energy consumption perspectives and the limitations that this perspective poses. The authors initially categorize the information parameters of the VHO processes, which vary from network-related context information (such as network load) to physical layer's indicators (RSRQ/RSRP); at a second stage, they attempt to link the energy consumption with the signaling- (and as a result delay-) overhead perspective; according to the authors, the majority of the literature efforts primarily focus on the type of parameters to aggregate for the context building, rather than the actual cost of these choices, both from the energy consumption, as well as the signaling overhead perspective.

Several other existing surveys attempt to present a unifying perspective with regard to HO mechanisms. Rao et al. in [37] deal with the network selection concept as a perspective approach to the always best connected and served paradigm in heterogeneous wireless environment. From the origin point of view, they classify them in four categories: network-related criteria, terminal-related, service-related and finally, user-related. In addition, in [38]-[40], several efforts are described, which aim to improve the selection mechanisms, which support heterogeneous RATs. In principle, all mechanisms combine parameters like Received Signal Strength (RSS), bandwidth, mobility, power consumption of the UE, security, monetary cost and user preferences. In all the above cases, the researchers are using for the most advanced schemes a number of parameters. However, very rarely they clearly state how this information is collected and from which network entities. Such information is necessary because the hypothesis that a value (e.g., the location of terminal) can be collected may require extensive signaling exchange among the network components.

Lately, numerous novel efforts found in journal papers, which address the handover mechanism, traffic steering and RAT selection procedures for future networks have been published [41]-[48].

In [41], the authors focus on the handover delay challenges, from the handover security and user authentication perspective. Ultra Dense deployments may result in frequent handovers, which may subsequently introduce high delay overheads. They propose the Software Defined Networking (SDN) enabler as one of the most promising solutions; through its centralized control capability, user-dependent security context may be exchanged between related access points and enable delay-constrained 5G communications. The context is shared between nodes and APs based on UE path prediction. A redesign on the "intra-macrocell" handover procedure is described in [42], focusing on the control/user plane split HetNets of future systems; the handover optimization is realized by predicting the received signal quality of the UE, triggering as a result the handover decision in a more efficient way. The authors focus on the challenges of the handover between macro cell at high speed scenarios (railway, highway, etc.). The respective evaluation shows that by predicting the forthcoming UE measurement reports, the handover execution takes place in advance and the handover performance is enhanced.

Similarly, in [43], the authors also focus on the -control and user plane separated- future HetNets and more specifically, on the signaling latency reduction in cases of macro cell base station fail-over periods. The proposed solution is based on a small cell controller scheme for controlling and managing small cells boundaries in a clustered fashion, during the corresponding macrocell's fail-over period. The evaluation of the proposed scheme on the UE side demonstrates reduced signaling latency, particularly for high user velocities; however, at the same time, the data delivery latency increases comparing to the legacy scheme; the authors conclude that the application of the proposed scheme can be selected on the specific signaling and data delivery latency requirements of each use case.

The RAT selection and handover procedure have been also studied from lower levels' perspectives as well. The very recent work in [44] focuses on the high frequency bands above 6 GHz, which will provide considerably larger bandwidths than the legacy systems; the challenge that the authors identify relates to the modifications that need to take place in the design of some key functions, such as the handover in order to support future deployments. They propose a novel frame structure, flexible and scalable to support various numbers of beams/antennas, users, or traffic conditions. The evaluation that was conducted involved static, as well as high velocity UEs; the authors conclude that the proposed enabler succeeds at satisfying all the throughput and delay requirements of the forthcoming 5G and beyond use cases.

Handover management in ultra dense heterogeneous small cell networks is studied in [45], focusing on the cell edge users. The authors describe an architecture comprising a cloud radio access network (C-RAN), as well as base band unit (BBU) pools, in which resource management and control capabilities are co-located, such as handover decision function and admission control. The proposed handover is realized between the BBU pools. The evaluation of the proposed scheme showed that the capacity of the small cells is increased, without increasing however the QoS of the users as well. In [46], the authors outline the main challenges that come with the UDNs. Among numerous challenges, such interference mitigation, backhaul issues and energy consumption, the authors tackle the mobility and handover

challenge as well. Among the enabling technologies they propose is cell and receiver virtualization, self-backhaul solutions and user-centric control of user information to minimize signaling.

The challenges posed by the UDNs are discussed also in [47]. The authors present an overview of the existing solutions related to control and data plane separation architecture (SARC), which they consider as one of they key enablers of the UDN use case, while they focus on the coordinated multipoint (CoMP) and Device-to-Device (D2D) communications. According to the authors, SARC can optimize considerably the handover mechanism and minimize the signaling overheads that were imposed until now: for example, users with no active data sessions will not have to be handed over. Similarly, in [48], the authors study the 5G UDNs challenges and propose a novel architecture that absorbs the Machine Type Communication (MTC) high and unpredictable traffic via home eNBs, allowing them to significantly reduce congestion and overloading of radio access and core networks. The main goals that the authors address are a stronger separation between MTC and Human Type Communication (HTC), closed access femto cells , -which will be only available to those machines that belong to a given closed subscribed group-, as well as coverage extension. In relation to the handover mechanism, the primary claim of this work relates to the X2 interface, serving a low-latency interface to exchange data traffic for time critical events of MTC traffic. Via extensive evaluation scenarios involving MTC UDNs, the authors demonstrate considerable gains in terms of latency in the handover procedure, energy consumption, capacity and scalability.

In relation to the ANDSF network entity, which will play a major role in the future Dense HetNets –as discussed earlier-, we must highlight the fact that the only literature proposals attempting to address the 5G RAT selection topic taking into account the ANSDF features of the latest 3GPP release are [49], [50] and [51]. In [49] the authors provide detailed insights with regard to the means that the context information is acquired and by describing how ANDSF is utilized for retrieving the desired information items, i.e., the BS transmitted power, the cell traffic load and the user's spectral efficiency. Nevertheless, the authors do not address the issue of the frequent handover process trigger as they solely rely on the A2 handover event (RSRQ threshold), while in addition no reference is being made to the velocity of the UE, an essential parameter need to be taken into account for also avoiding unnecessary handover triggers and ping-pong effects. Finally, no discrimination is being made depending on the type of the traffic flow, implying that all traffic flows –no matter their specific QoS requirements- are handled the same way. On the other hand in [50], the authors choose to use the ANDSF to facilitate the discovery of non-3GPP access networks; however, the context information that may be residing in ANDSF is not claimed to be taken into account for the optimization of the handover procedure, as the authors select to follow the existing handover techniques. Finally, in [51] the authors propose an energy-efficient handover mechanism, based on the ANDSF, which attempts to minimize the overall power consumption, maintaining a minimum QoS for the active sessions.

In most cases solutions target either handovers for macro-femto cells or vertical handovers among different RATs as separate approaches. In this paper, we attempt to overcome such discriminations and approach the problem from a unified perspective; this is further supported by clearly indicating how the information required for our solution is collected and from which network entities.

Last but not least, an interesting observation is that no scheme available in the literature seems to take into account the load of the backbone, i.e. the backhaul load, a parameter, which is of paramount importance as well; investigating only the load of the network edge may be misleading; in the proposed scheme we take this parameter into account in our RAT evaluation as we show in the next section.

As far as the core evaluation mechanism is concerned, i.e. the mechanism, which receives the input context information, processes, evaluates it and generates the handover decision info, diverse solutions have been proposed and used. For our solution we have also chosen to use a Fuzzy Logic (FL) scheme to support the decision making process. In [38], a significant number of papers is analyzed. The authors make a very detailed categorization of the available approaches of the different groups –according to the core evaluation mechanism presented-: RSS-based, QoS-based, Decision Function based (e.g., Utility Functions, Cost Functions, Multiple Attribute Decision Making –MADM schemes), as well as Intelligence based (e.g., fuzzy logic and neural networks schemes). As it can be rationally hypothesized, all categories have their advantages and disadvantages.

In general, RSS-based schemes primarily demonstrate reduced handover failures and ping-pong effects, while in some schemes throughput improvements are observed as well; on the contrary, their limitations seem to relate to increased packet loss, increased signaling and -as a result-, increased handover latencies, which makes them unsuitable for real-time applications. QoS-based handover schemes demonstrate higher throughputs, less packet loss and overall assured

QoS, while they are not applicable for high-speeds and are sometimes prone to ping-pong effects. Decision Function based schemes are cost and energy effective and are less likely to experience ping-pong effects, however, they are also unsuitable for real-time applications and they sometimes pose excessive burden on the network, which leads to degraded QoS. Network Intelligence based schemes achieve reduced delays and packet loss, while they are efficient in terms of handover success as well. On the contrary, they are slow at the training and learning phase, their complexity is considered high, while their centralized control is also considered a drawback. The Context-based schemes in general (such as the proposed one) are adaptable to different network technologies, demonstrate optimized blocking rates and handover latencies, optimized throughput, while one of their main advantages is also their often fully distributed nature. However, the most crucial drawback of context-based mechanisms is the increased communication overhead required, their instability for variable speeds and the deployment issues in the real world.

Overall, fuzzy logic context-based schemes are a good choice for the problem we study for a number of reasons: They have better performance compared to simple RSS or QoS-based schemes. Their behavior is better, compared to decision function based schemes (e.g., MADM), when the monitored parameters contain imprecise data. Also, fuzzy logic mechanisms can be used as a supporting-tool for context based schemes. Finally, today's smartphones can easily support fuzzy logic schemes based on their capabilities (e.g., processing power, memory etc.).

Many authors have proposed FL Inference Systems. Indicatively, Xia et al. [52] propose a scheme taking into consideration the actual RSS, as well a predicted RSS, and they combine it with the speed of the UE in order to determine if a handover should be made or not. Moreover, they estimate the suitability of a RAT for handover, taking as input the current RSS, the estimated RSS, as well as the available bandwidth. In [53], FL is also used for estimating the output suitability of a network based on the inputs of the environment (bandwidth, delay, charging, power consumption). In addition, Ma and Liao use GPS, in order to adapt the monitoring rate of the aforementioned values.

Despite the numerous efforts using fuzzy logic, however, to the best of our knowledge, there is no work related to 5G cellular systems. That is because mostly, existing fuzzy logic related research proposals focus on UMTS and WLAN networks. One of the most important gaps in relation to these efforts, however, is that none of the previous papers provides a detailed description on the means, via which the contextual information is being collected from the involved components. This is crucial since the acquisition of multiple parameters may place a heavy burden to the network infrastructure. The proposed scheme does provide the specific context sources inside the network environment; more importantly, this has been done fully in line with the current 3GPP directions and trends of next releases. Finally, previous authors do not base their work on a realistic business case; -thus- they have to take into consideration too many or too few input parameters for their mechanisms.

In the following section we provide a thorough description of the proposed algorithm, we describe our mechanism based on a specific business case, we provide viable solutions on how the contextual parameters can be collected without affecting the network infrastructure and we focus on a typical 5G cellular system deployment. In addition to the aforementioned, we provide extensive simulation results, which prove the validity and viability of our solution.

III. COMPASS: THE CONTEXT - AWARE RAT SELECTION 5G MECHANISM

This section focuses on the comprehensive presentation of the proposed mechanism. The section is divided into four sub-sections; the first one provides some initial insights in terms of inputs, the decision making process and outputs; next, the algorithm of the solution is illustrated and each step is discussed in a thorough way; the third sub-section focuses on the fuzzy-logic component of COmpAsS, as well as the applied rules; finally, the network architecture perspective is presented, in which the proposed mechanism is evaluated in terms of required interfaces, network entities, message types, etc.

A. The network architecture perspective

The mechanism –as already mentioned earlier- is user-oriented, i.e. deployed on the UE side, rather than one of the main LTE network entities. Nevertheless, its functionality is not completely independent as the architecture of the network is directly influenced in the sense that the context information, which is required to be aggregated by

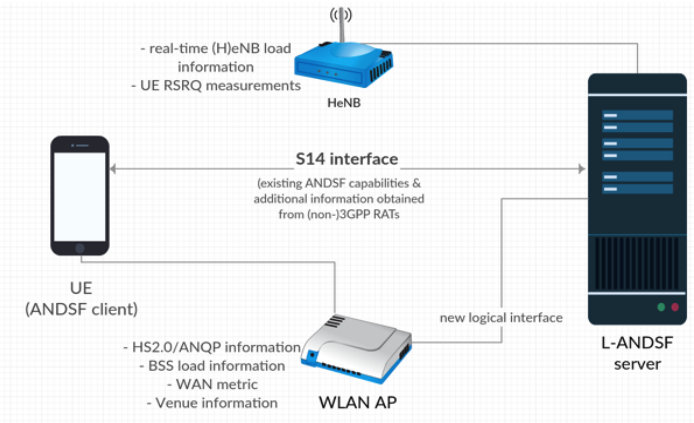

Figure 1 Context information acqusition via the respective network

the UE, is available in specific network components. Moreover, the mechanism is directly associated with the network policies, in the sense that the UE-oriented decisions are being forwarded to the central decision-making entity in the network core, which will make the final assessment. As a result, some minor adaptations need to be realized in order to enable the required context information acquisition. In this subsection, we address the requirements in terms of the data sources, message types, as well as interfaces, which are required to support the proposed mechanism.

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

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

More specifically:

- The *RSRQ value* is already part of the UE measurements report used in LTE for evaluating the quality of the signal of the neighbor base stations. Similarly for Wi-Fi, RSS metric is already included in the existing IEEE 802.11 reporting metrics, even for the end-user devices.
- As indicated in [55] the *mobility state* of the UE (high-mobility state, medium, etc.) is considered and is sent via the system information broadcast from the serving cell.
- The *traffic flow type*, mapped to the respective sensitivity to latency for each application/service type executed by the UE: the different application categories and respective flow QoS requirements are extracted by the UE connection manager, from the well-established port numbers of the applications/services.
- The *traffic load* of the base stations and the *backhaul load* of the network: In the proposed mechanism, the UEs collect information from a local instance of ANDSF (Local ANDSF - L-ANDSF) as suggested in [26] about the policies of the operator for accessing WiFis in the area, as well as information from Hotspot 2.0 protocols to evaluate the status of WiFi APs (e.g., number of users associated to the AP, the load of the backhaul link of the AP etc.). The main functionality and role of the ANDSF has already been discussed. We extend this concept by assuming that local ANDSF entities (L-ANDSF) contain similar information (i.e., number of associated users, load of the network link, etc.) for every (H)eNB in a specific area (Figure 1). This distributed model radically decreases the information exchange delays between the nodes in a limited area, comparing to a scenario, in which one central ANDSF entity of the operator serves thousands or millions of devices. This requires appropriate logical interfaces from the (H)eNBs to the ANDSF. Last but not least, the nodes update L-ANDSF in a coarse manner (e.g., Load is Low, Medium or High) only when thresholds are violated, so as to further minimize the signaling overhead in the network.

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

B. Overview of the proposed solution

The proposed RAT selection mechanism presented in this work aims at enabling the UEs to identify in an intelligent way the most suitable RAT to associate with in a specific urban area, where a cellular operator (with deployed macro, pico or femto cells) co-exists with Wireless Internet Service Providers (WISPs) with whom it has roaming agreements as suggested in [33]. The mechanism is applicable for users of 5G smartphones that support a number of RATs.

The framework is using pre-defined, customizable and fine-tuned rules for all the possible combinations of the different aforementioned scheme's inputs. The rules that are applied are policies, based on objective network parameters, KPIs and general principles, derived from the state of the art of the domain, as it was presented in the previous section. More specifically, according to these rules/policies, a RAT, which is characterized by low (backhaul) load and high RSS/RSRQ, is advantageous for the UE choice. In addition, the higher the sensitivity to

latencies (traffic flow type input type), the higher impact the mobility metric has on the *Suitability*; high mobility UEs are preferably placed in larger cells to avoid unnecessary handovers and/or ping-pong effects. Using Fuzzy Logic Controllers (FLC) each UE evaluates the available RATs and identifies the most suitable one, which optimizes the Quality of Service (in terms of pre-defined KPIs) for each application (or type of traffic); afterwards it performs a session initiation or a per flow-handover using existing 3GPP mechanism described in the introductory section. The KPI, which is utilized to describe the selection prioritization among heterogeneous cells and access technologies is denoted as *Suitability* in our scheme (Figure 2).

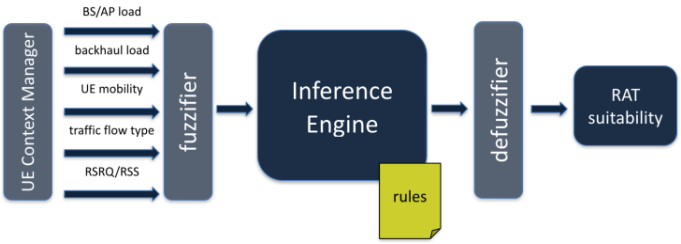

Figure 2 Fuzzy Logic Controller for the extraction of the RAT *Suitability* metric

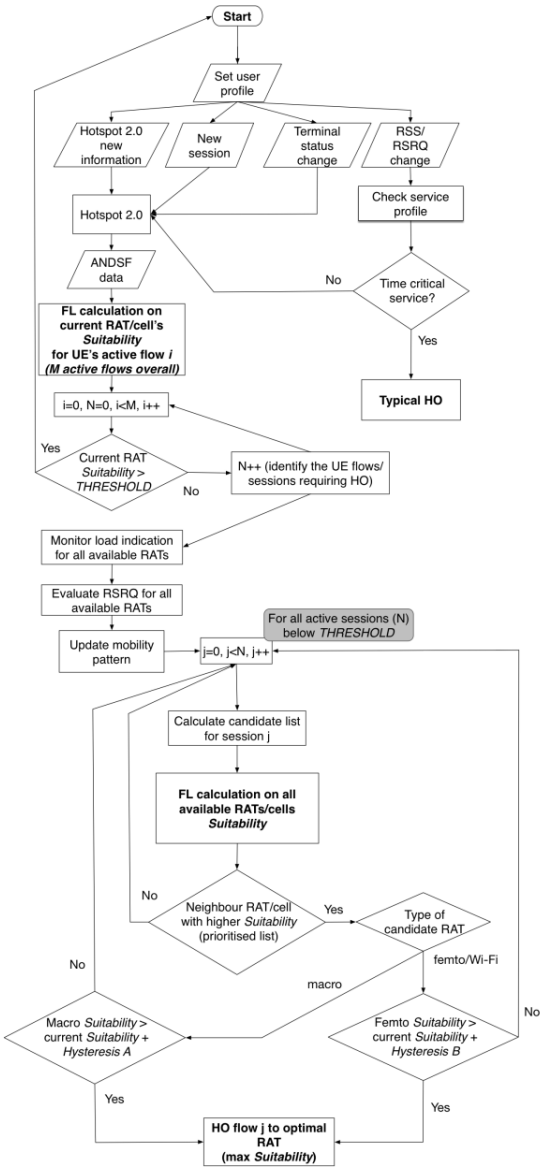

Figure 3 The algorithm of the solution

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

Table I N_F x N_R *Suitability* Calculation example for a UE with 4 active IP flows

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

The proposed scheme's decision making process selects for each one of these active flows the RAT, for which *Suitability* is maximized; afterwards, the UE makes a handover request to the respective (H)eNB or AP in order to transfer the flow to the optimal access technology. The process is running both on a pre-defined time interval basis, as well as upon pre-defined trigger events, which are described in detail in the following sub-section. If for any reason, the handover to the highest-ranking RAT is not possible, the 2nd choice in the *Suitability* list is selected, etc.

C. *Description of the algorithm*

In this section we present the algorithm we designed for the optimal RAT selection per application/session (Figure 3). As already mentioned in the beginning of this section, the alpha version of the scheme was evaluated and presented previously [23],[24]. A detailed

discussion has been provided at the last part of the introduction section, which describes in detail all the novel features and algorithm parts of this extended work.

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

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

The values of the *Threshold* and the *Hysteresis* may be configured according to the specific needs of a particular network environment by the network administrator before the mechanism is deployed on the UE. An extension of this feature that could also be accommodated in the future is the enablement of an automatic adaptation of the two control parameters, by defining the different possible "states" of the network and the respective *Threshold-Hysteresis* configuration for each one of these states. For example, for a more dense, -in terms of network deployment-environment, the solution performs better for higher *Threshold* and *Hysteresis* values. It should be also noted that the network "state" sensing is already enabled via the available context information, which is being aggregated by the UE. The algorithm is described thereinafter step by step: initially, the user defines either on a per session basis (e.g., HTTP traffic to be handled only by free WiFi) or collectively (e.g., use always the RAT that minimize the energy consumption) his preferences (i.e., "user profile" in the algorithm flowchart). For the user profile generation numerous solutions have been proposed, i.e. either manually or automatically using data analytics solutions; for our algorithm, a solution that suggests the profile creation in an automated manner is selected. The mechanism algorithm may be triggered only if there is at least one session active in the UE. Thus, as long as there is at least one session, pre-defined events trigger the algorithm initiation, i.e.: new information from Hotspot 2.0 is received (e.g., a WiFi AP is now unloaded), a new session is initiated on the UE side, a significant change in the terminal status (e.g. battery level is falling below a certain threshold) or a significant change in the monitored RSS/RSRQ values is identified, etc.

By the time the process is triggered, and only if there is no time critical service to be served, (-in which case RSS/RSRQ has fallen below a threshold and a typical HO must urgently take place-), the UE proceeds to the information aggregation phase, which is being supported by Hotspot 2.0 and ANDSF servers, in relation to all the available neighbor RATs (3GPP or non-3GPP) and cell layers (macro, pico, femto cells, etc.), without any direct association with them. This information relates to the available 3GPP or non-3GPP cells' and APs' load, number of associated UEs, quality of received signal, etc. As already discussed in Section II, in 3GPP Rel-12 and beyond, the ANDSF enhancement creates a sufficient RAT context source. This is achieved by incorporating additional information items, better granularity for the existing for traffic steering conditions [54], as well as integrating information from Hotspot 2.0. This information is updated on the UE side, triggered on a per trigger-event basis.

Having aggregated the updated context information from the aforementioned sources, the UE performs an updated calculation on the *Suitability* of the currently associated RAT/cell, i.e., whether it is over the preset *Threshold* value. If yes, then no further calculation or signaling is required and then algorithm reverts to the starting point.

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

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

D. The Fuzzy Logic modeling of the solution and the rules applied for decision making

As already discussed earlier in sub-section IIIC, as well as illustrated in Figure 2, we identify five input context parameter types, which are considered as of paramount significance, for the evaluation of the *Suitability* of a candidate RAT/cell for handing over an active UE session.

The acquired information is processed using a Fuzzy Logic Controller (FLC), which is the core decision-making mechanism of the proposed scheme. Fuzzy logic is an ideal tool for dealing with uncertainty cases, when the inputs are rough estimated values. Furthermore, fuzzy logic handles the curse of dimensionality by using generic input and output states. Finally, the fuzzy logic is a tool for handling multi-variable problems, where a joint correlation analysis of several inputs is required [20].

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

The proposed scheme uses three FLCs, each one for every RAT type, i.e., macro cells (LTE, GSM, etc.), femto cells and WiFi APs. Every time that the algorithm is being triggered, all the available base stations and APs are being evaluated. The inputs of every fuzzy reasoner are the RSRQ/RSS, the service sensitivity to latency, the load of the (H)eNB or AP, the backhaul load of the corresponding (H)eNB/AP, as well as the UE mobility status. Each of the

inputs is being fuzzified to low, medium, and high membership functions (MFs). All the inputs apart of the service sensitivity to latency are being fuzzified using triangular and trapezoidal MFs for exploiting the fact that each state is in general well defined, and has zero values for each state. On the other hand, for the service sensitivity to latency, we have used Gaussian input MFs, for highlighting the fact that all the states have non-zero values; in other words, even for the services that are considered as non-latency sensitive the user still has (loose) latency requirements.
The inputs are being combined in the FIS as defined by the rule sets that have been defined in the following format:

Equation 1 Fuzzy Inference System's Rule Example

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

In this case for every of the 3 fuzzy reasoning modules we have defined 243 rules to cover all the potential input combinations, resulting in 729 rules overall. As also discussed earlier, these rules are subject to adaptations, according to the administrator's policies.
In order to elaborate on the input parameters and the fuzzification rules and provide a comprehensive picture of the way they influence our system, we illustrate how the input context parameters are fuzzified inside the FLC, as well as how the output is defuzzified as RAT Suitability, by providing a number of indicative 3D surfaces (Figure 4-Figure 11) that have resulted from the aforementioned rule sets. The *Suitability* value ranges from 0,0 to 1,0 (0-100% respectively) in the vertical axis. It must be noted, in addition, that each plot illustrates only two out of the overall five inputs at a time -for visualization reasons-.

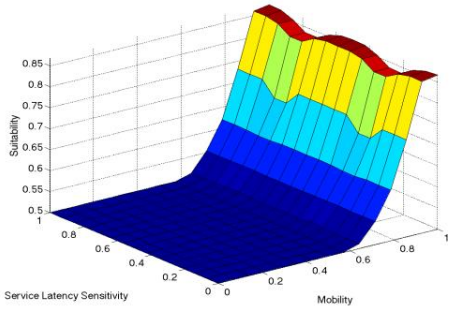

Figure 4 $Suitability_{macro}$ = f (Latency sensitivity, Mobility)

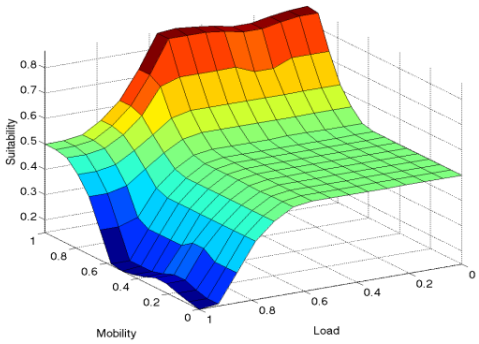

Figure 5 $Suitability_{macro}$ = f (Mobility, Load)

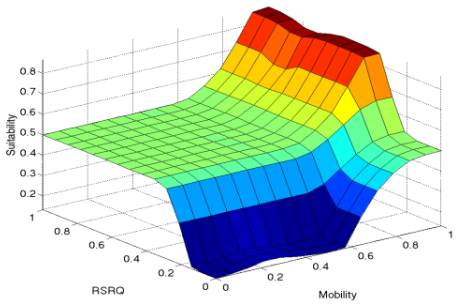

Figure 6 $Suitability_{macro}$ = f (RSRQ, Mobility)

As illustrated above (Figure 4), the *Suitability* of a macro cell ranges from 50% up to 85% (0.5 and 0.85 respectively on the z axis) for varying mobility and service latency. Specifically, it can be seen that a macro cell becomes radically more suitable when the mobility of the UE increases. This means that the proposed scheme prioritizes the placement

of high velocity UEs to big cells, due to their advantage over the small cells in relation to the considerably higher time frame a fast UE may remain within a macro cell's boundaries. Lower time within a small cells boundaries leads to increased delays that are introduced by the connection establishment procedures upon a handover. The traffic type's influence on the Suitability is almost negligible, when comparing to the mobility. Similarly with Figure 4, the UE mobility has the identical impact with the previous example in Figure 5. Contrary to the previous case, in which the mobility played the dominant role, this time the 2nd parameter, -the traffic load of the base station-, plays a major role, causing the Suitability to decrease considerably when the load increases, and vice versa. In addition, it can be observed that when the load is less than 60% (meaning that most probably the base station is capable of serving a new UE session), only mobility influences the final outcome. Figure 6 shows how RSRQ influences the Suitability of a macro cell as well. Once more, the mobility's impact is explained in accordance with the two previous surfaces. One detail worth discussing is that all RSRQ values above 20% (of the max RSRQ) are acceptable; however, below that threshold the cell becomes inappropriate due to very poor connection quality, and the Suitability falls instantly for any type of UE mobility.

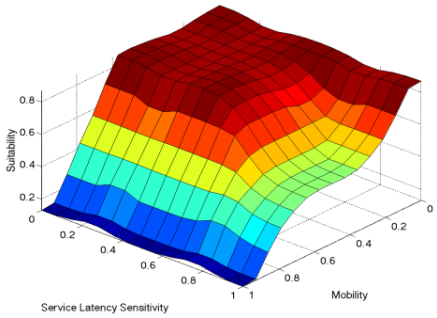

Figure 7 $\text{Suitability}_{\text{femto}}$ = f (Latency sensitivity, Mobility)

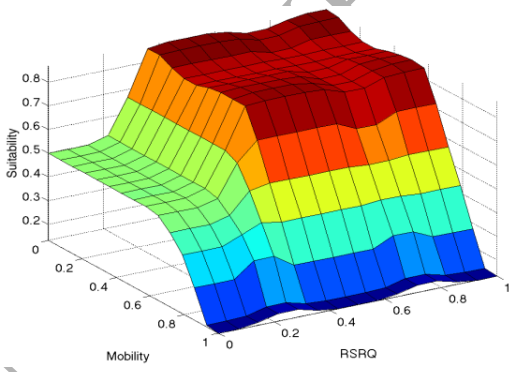

Figure 8 $\text{Suitability}_{\text{femto}}$ = f (Mobility, RSRQ)

Contrary to the macro cell's selection criteria, the higher the mobility of the UE is, the lower a femto cell's *Suitability* becomes. The reason behind this radical decrease –as it is illustrated in Figure 7- is due to the limited time frame a UE may remain within a small cell's boundaries, leading to increased delays that are introduced by the connection establishment procedures upon consecutive handovers. In other words, the proposed scheme mostly selects to offload traffic to femto cells by priority to static or slowly moving UEs. Figure 8 shows the same pattern with the respective macro cell's surface (Figure 6). The main difference relates to the inversed mobility's impact, as also discussed earlier.

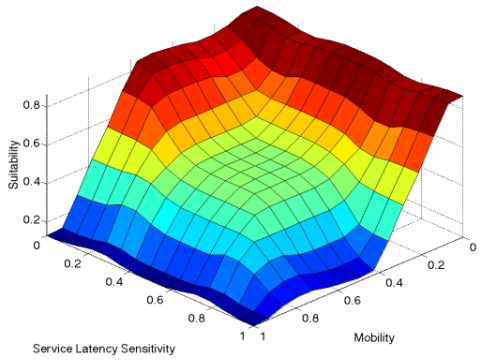

Figure 9 $Suitability_{WiFi}$ = f (Latency sensitivity, Mobility)

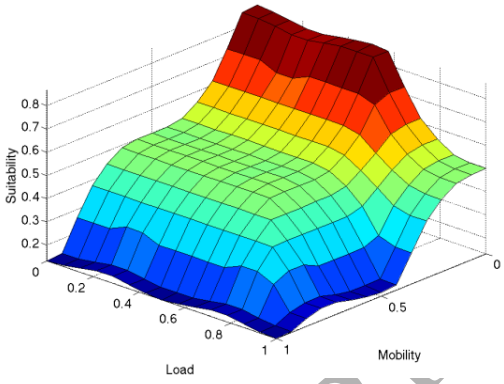

Figure 10 $Suitability_{WiFi}$ = f (Load, Mobility)

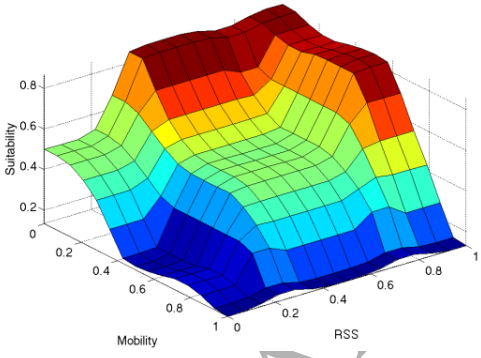

Figure 11 $Suitability_{WiFi}$ = f (Mobility, RSS)

The selection criteria of WiFi APs resembles the femto cells' due to their limited range and –as a result- the time a high mobility UE may remain within the cell's boundaries and avoid consecutive handovers or ping-pong effects. In addition, WiFi APs are suitable for serving services with high sensitivity to latency (Figure 9) due to the number of resources usually shared among fewer users, resulting in more effective resource scheduling among them. In accordance with the aforediscussed parameters and their impact on the respective RAT Suitability, Figure 10 and Figure 11 illustrate the Load – Mobility and Mobility – RSS impact on the WiFi AP's selection *Suitability*.

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

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

Monitored context parameter	***Impact on Suitability***	***Correlation with other input parameters***
BS/AP load	Inversely dependent (higher load, lower *Suitability*)	a) Directly linked to the backhaul load b) The higher the RSS/RSRQ the lower the impact of load on the *Suitability*

Backhaul load	Inversely dependent	a) Directly linked to the BS/AP load b) The higher the RSS/RSRQ the lower the impact of backhaul load on the *Suitability*
UE mobility	Inversely dependent for small cells (femto/Wi-Fi) Proportional for large cells (higher mobility, higher *Suitability*)	Linked with the traffic flow type: the more sensitive to latency, the higher the impact of the mobility on the *Suitability*
Traffic flow type/ Sensitivity to latency	Inversely dependent for small cells (femto/Wi-Fi)	a) Linked with the UE mobility (see above) b) Influenced by the RSRQ/RSS: the higher the RSS, the less *Suitability* is influenced by the Sensitivity to Latency
RSRQ/RSS	Proportional (higher RSRQ, higher *Suitability*)	N/A

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

E. Signaling-related issues

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

We assume that O_{output} equals zero, as the *Suitability* ouput report of the UE plays an identical role to the current 3GPP's *UE Measurements Report* consisting of the reported RSRQ values. As a result, input signaling cost is directly linked to the five input parameters (Equation 2):

Equation 2 Signaling Cost

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

U is the overall number of Users (UEs) in a particular network environment. *R* is the overall number of the available RATs, which are evaluated by the UE. S_X refers to the additional signaling payload of the respective input parameter. According to the earlier analysis (Sub-section III-A), three out of the five input parameters ($S_{RSRQ/RSS}$, $S_{mobility}$ and $S_{flowtype}$) actually correspond to zero overhead as they are already available according to the current 3GPP standards. The traffic load information both for the access, as well as the backhaul link are described in a coarse manner (low, medium, high); as a result, we append the additional required number of bits in the signaling messages.
The figure that follows (Figure 12) illustrates how the signaling increases for the legacy RSRQ/RSS-based mechanism and the proposed one in several 5G use cases. The evaluation has been made for one evaluation of all available RATs by all the UEs in an area. To further elaborate, we consider some of the most challenging 5G use cases, which follow the METIS project specifications [32]. According to these specs, each use case corresponds to a different number of a) UEs and b) RAT choices, in the specific network environment; virtual reality office, shopping mall, traffic jam, and stadium/open-air festival are some key examples, which demonstrate an increasing number of users and available RAT options in a specific area, ranging from a few UEs and RATs in an a area (Virtual Reality Office) to thousands (Stadium).

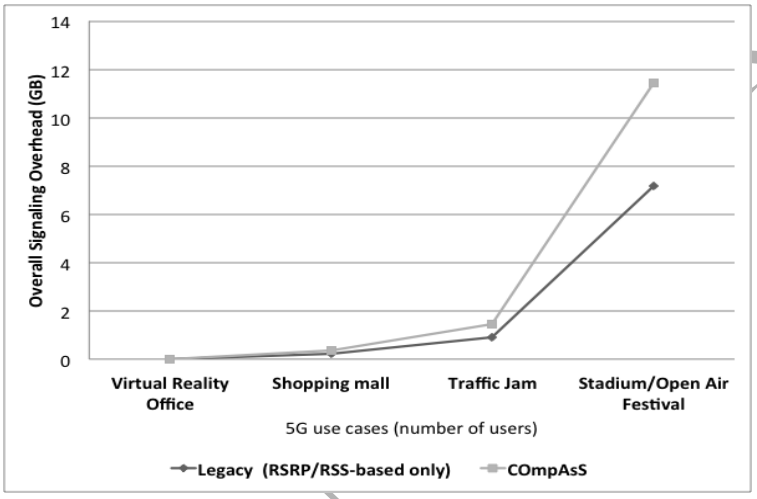

Figure 12 Signaling overhead for advanced context acquisition in diverse 5G use cases

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

IV. EXPERIMENTAL EVALUATION

In order to assess the validity and viability of our approach, we performed extensive experiments using the NS3 network simulator and -customized for the particular evaluation needs- Python modules. We considered a usage scenario from the METIS project [56], which we implemented in NS3 and evaluated the application of COmpAsS on complex RAT environments. Through the experiments we attempted to replicate –to the best possible extent and taking into account the simulator's limitation– a real life situation. Towards this end we conducted an extensive literature review that covered a large number of aspects like mobility speed, service usage patterns etc. We present our findings and configuration in the following.

A. Experimentation Scenarios and Setup

The main scenario considers one of the established 5G use cases, -as these were documented in [56]-, i.e., a large shopping mall with high density of customers and service staff (Figure 13). We selected this set-up, as a typical setting for a future extended rich communication environment, involving both "traditional" radio networks, as well as wireless sensor networks, where customers access mobile broadband communication services while they are directly addressed by personalized location-based services of the shopping environment. We evaluate this setting on the basis of 4 different scenarios, which we describe in detail below. Overall, the network deployment allows seamless handling of services across different domains, e.g. mobile/fixed network operators, real estate/shop owners and application providers. Based on this description, we use the NS3 and model a 3-floor shopping mall. Each floor's dimensions are 200x100m, containing 20 rooms/shops per floor, with several LTE Femto cell placed on each floors, depending on the scenario. Outside, two LTE eNBs are placed, 150m north and west of the mall respectively.

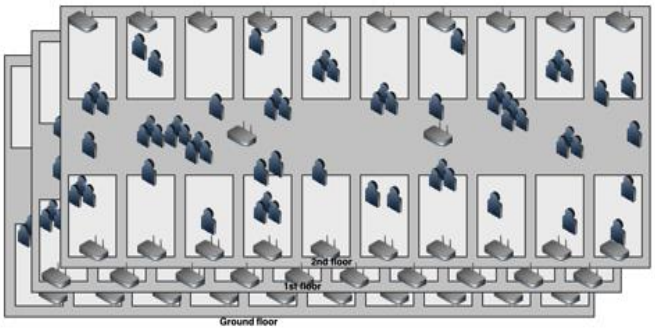

Figure 13 Simulation environment: Shopping mall comprised of 3 floors and 20 shops per floor

In order to evaluate the proposed framework, using LTE femto cells and macro cells is sufficient, as the rules that apply for Wi-Fi are almost identical with the ones applied for femto cells. In addition, the IP flow mobility between LTE and Wi-Fi networks is not available in the NS3 simulator that was selected. Our simulation scenarios are based on 3GPP Specifications [57] and [58]. In details, the transmission mode is SISO (Single Input Single Output) and the scheduler is the NS-3 implementation of the Proportional Fair MAC scheduler. We use the Hybrid Buildings Propagation Loss Model for path loss implemented in NS3 with Internal Wall Loss at 10.0 db Shadow, Sigma Indoor at 10.0 db. The network node configuration appears in Table III. Services are implemented using NS3's UDP client-server application model and the desired data rates are achieved through configuration of the packet size and the inter-packet interval parameters. The service schedule for every user is pseudo-randomly generated at the beginning.

Table III NS3 Simulations' Configuration

NS3 Network Node	**Tx Power (dBm)[57]**	**Downlink (DL) Earfcn (MHz) [57]**	**Bandwidth (RBs) [57][58]**	**Antenna Type [57]**
Macro cell	35	2120	50 (10 MHz)	Parabolic, 15 dBi
Femto cell	20	2120	15 (3 MHz)	Isotropic
UE	20	-	-	Isotropic
Other parameters				
Number of eNBs	2			
Number of HeNBs	50 (max.)			
Number of UEs	50			
Simulation time	225 s			
Time unit	0.1 s			
Transmission mode	SISO (Single Input – Single Output)			

B. Experimentation Methodology

As already presented in detail in Section IIIC, the proposed framework's algorithm uses two parameters, i.e. *Suitability Threshold* and *Hysteresis*. Different parameter values may alter radically COmpAsS's responsiveness and functionality, primarily in terms of triggering events frequency. Different network "states" (e.g., denser or scarcer deployments) would require different configurations of these two control parameters. Towards this fine-tuning process hence, in the first two scenarios, we incorporate in our experimentation a range of values, both for *Threshold* and

Hysteresis. Overall, the evaluation of COmpAsS moves along 4 axes-scenarios, each one of which focuses on a different varying parameter of the experiment's setup, in order to simulate -in the most realistic extent possible- all the radio conditions and network "states" that the proposed framework may encounter. In the following table, we present in more detail the 4 different scenarios.

Table IV Experimentation scenarios

Scenario #	**Scenario Parameter**	**Value range**	**Number of experiments**	
			COmpAsS	**A2A4-RSRQ**
1	*Suitability threshold*	[0.99, 0.9, 0.7, 0.5, 0.1]	75 experiments (15 different executions per threshold value)	15 experiments (*Suitability threshold* does not apply)
2	*Suitability margin*	[0.3, 0.2, 0.1, 0.01, 0.001]	75 experiments	15 experiments (*Suitability threshold* does not apply)
3	*Deployment density (number of HeNBs)*	[2, 5, 10, 20, 50]	75 experiments	75 experiments
4	*Network load (number of traffic bearers/UE)*	[1, 2, 3, 5, 10]	75 experiments	75 experiments

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

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

Our mechanism is juxtaposed against the well-established handover algorithm A2A4 RSRQ [59]. In addition to the aforementioned 300 experiments, another 180 experiments were run using the A2A4 RSRQ algorithm, resulting overall in 480 experiments, providing, thus, an extensive set of results for being able to accurately assess the validity and viability of the proposed scheme. A comprehensive analysis and discussion on these results will be presented in section C that follows. The Key Performance Indicators (KPIs) that are used for the evaluation follow:

- *Average Uplink (UL) Throughput:* The average UL throughput throughout the simulation duration
- *Downlink (DL) Throughput*: The average DL throughput throughout the simulation duration
- *Average Uplink (UL) Delay:* The average UL delay throughout the simulation duration
- *Downlink (DL) Delay*: The average DL delay throughout the simulation duration

In order to evaluate the above KPIs, we deploy randomly among the simulation UEs 4 different service categories, corresponding to 4 different traffic models, as well (see TABLE V). The simulation UEs are running the respective assigned services throughout the whole simulation duration. For each one of the service/traffic flow categories, different KPIs are applied, according to the particular flow QoS requirements.

TABLE V: Service/application parameters used in simulation and respective KPIs applied

Deployed Service Types, respective traffic models and KPIs		
Type	**Traffic model**	**QoS-related KPIs assessed**
1. Conversational voice	Average call duration is 1.8'[60] Average rate is 12.65 kbps[61][62]	UL/DL delay
2. Conversational video	Average video duration is 4.12' [65] Average DL speed 443 kbps [64] Average size for 480p video is 250MB per hour in [66]	UL/DL delay UL/DL throughput
3. Real-time gaming Data (non latency-tolerant)	Average packet size is 50kB Inter-packet time interval is 50 ms [67]	UL/DL delay

4. Non-GBR services (e-mail, chat, FTP, file sharing, etc.)	Average session for file download is 9.8s for 3MBs file [63]	UL/DL throughput

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

C. ***Results***

In this section we present the simulation results, which are used to evaluate the performance of the proposed mechanism. The baseline scheme that has been selected, i.e. the A2A4 algorithm, is a well-established LTE handover algorithm, which relies its decision making upon the RSRQ metric, as this is reported by the UEs. We divide the section per each one of the 4 scenarios, in four parts.

I. Scenario 1: Varying Suitability threshold

In our first evaluation scenario the *Suitability threshold* ranges between two extreme values: 0.1 and 0.99; taking into consideration that in the proposed scheme, context evaluation and *Suitability* calculation procedures are performed only when the current RAT's *Suitability* has fallen below the *threshold*, the behavior of COmpAsS varies significantly, primarily as far as the triggering frequency of the mechanism is concerned. The extreme value range has been selected on purpose, targeting to demonstrate how the algorithm responds under any possible configuration. The first set of results (Figure 14-Figure 17) depicts the throughput and delay KPIs both for the uplink and the downlink.

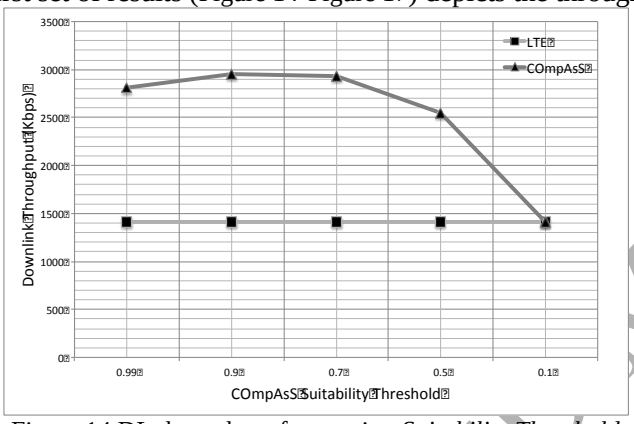

Figure 14 DL throughput for varying *Suitability Threshold*

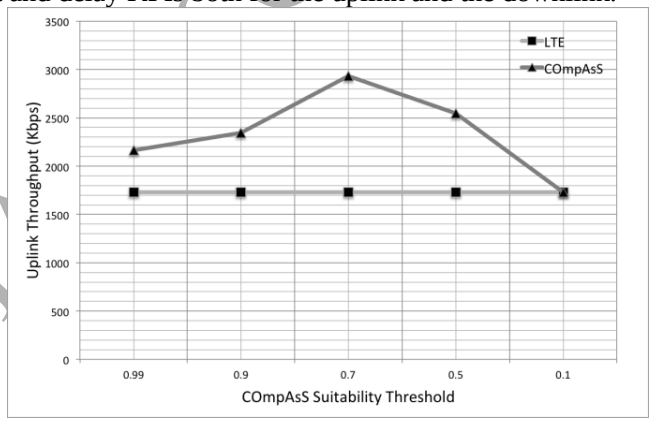

Figure 15 UL throughput for varying *Suitability Threshold*

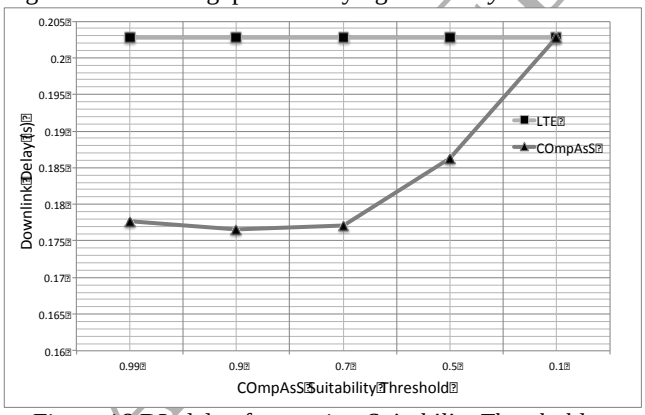

Figure 16 DL delay for varying *Suitability Threshold*

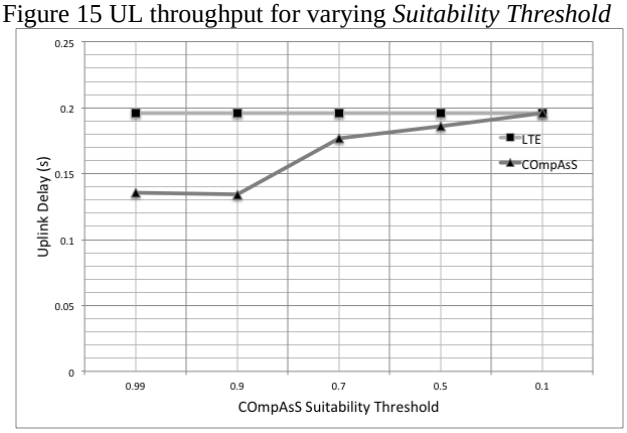

Figure 17 UL delay for varying *Suitability Threshold*

Studying the throughput graphs (Figure 14, Figure 15) reveals how COmpAsS outperforms the conventional LTE algorithm for all threshold steps, both in the downlink and the uplink respectively. In some cases, COmpAsS's performance is higher by more than 100%. Similarly, in the delay graphs (Figure 16, Figure 17), particularly for high threshold values, in the downlink, the proposed scheme's measured average delay is reduced by 10%, while in the uplink up to 25% when comparing to the A2A4 RSRQ.

High attention should be paid to the trade-off resulting from the *Suitability threshold* value, when experimenting with extreme high and extreme low values. Figure 15 demonstrates such a case, in which the optimized configuration results when *threshold* equals an intermediate value, i.e. 0,7. Extremely high values result in excessive number of algorithm triggers, which –although result in the optimal RAT selection at any point of time- cause signaling overhead and calculation latencies. In other words, for every network setting and environment, there is an optimal value for the *threshold*, for which the network metrics are the highest; this does not always corresponds to the highest *threshold* value. All in all, the configuration of the system in terms of the *threshold* relies each time upon the specific use case, the administrator's targets and depends heavily on the specific environment's conditions and QoS requirements.
Nevertheless, it must be highlighted that in all graphs (throughput and delay, both DL and UL), the two mechanisms' performance converges as the *threshold* decreases, while it becomes identical for the lowest *threshold* value (i.e., 0.1). This is explained by the fact that, the lower the *Suitability threshold* value, the fewer times COmpAsS is triggered and thus, the less effective it is throughout the overall simulation duration. When the value reaches 0.1, the system responds as if the algorithm is practically no longer active.
In accordance with the above results, we select the value of 0.7 for the next scenarios of the experimentation; using the specific value, the throughput optimization is highest, while the delay remains considerably lower than the baseline, both in uplink and downlink.

II. Scenario 2: Varying Suitability hysteresis (margin)

The second evaluation perspective illustrates the simulation outcomes in relation to the *Suitability hysteresis*, in a similar pattern with the first scenario; the *hysteresis* ranges between two limit values: 0.3 and 0.001. As it can be inferred, the higher the *hysteresis (*i.e. the difference between the current RAT's and the candidate target RAT's *Suitability* value), the "stricter" requirements of COmpAsS in terms of *Suitability* of the successor RAT for handover. On the contrary, the lowest *hysteresis* value, i.e. 0.001, implies that a handover is realized by the time a RAT with higher *Suitability* is detected, without actual margin being taken into account. The following figures (Figure 18-Figure 21) illustrate the results of the 2nd experimentation round:

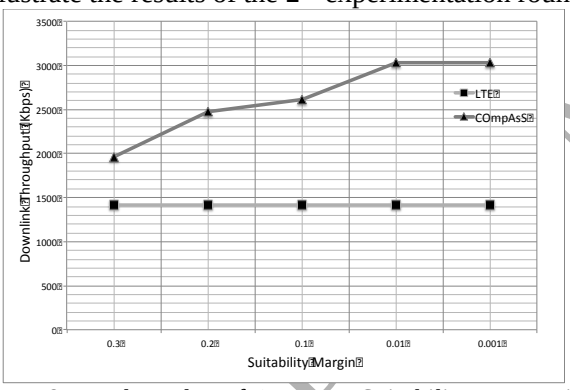

Figure 18 DL Throughput for varying *Suitability Hysteresis*

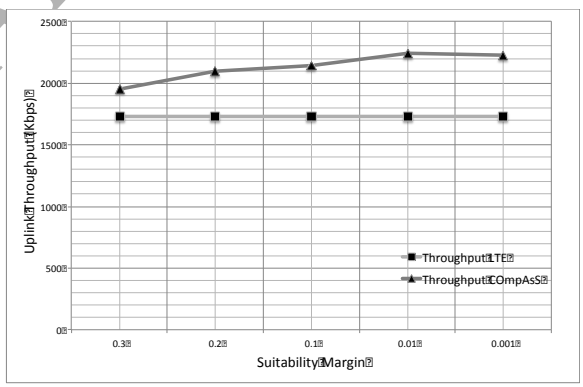

Figure 19 UL Throughput for varying *Suitability Hysteresis*

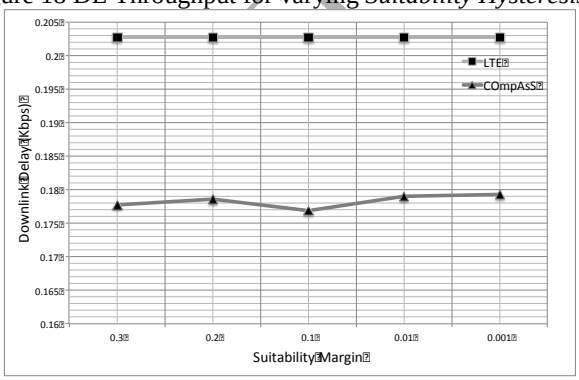

Figure 20 DL Delay for varying *Suitability Hysteresis*

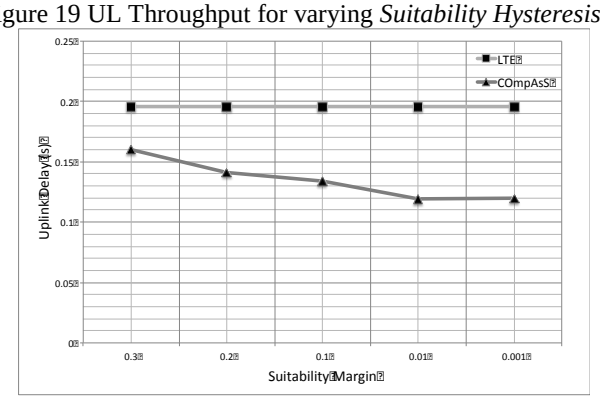

Figure 21 UL Delay for varying *Suitability Hysteresis*

It is clear that the proposed scheme's results showcases a clear superiority, both in terms of throughput and delay. The throughput is constantly increasing as the *hysteresis* is decreased, both for the uplink and the downlink. This is

expected in the sense that the lower the *hysteresis*, the more handovers are realized according to COmpAsS's decision making and RAT selection. As far as the delay KPI is concerned, the proposed scheme outperforms the LTE legacy algorithm for all hysteresis values. In the downlink (Figure 20) the delay is not influenced by the hysteresis, while in the uplink (Figure 21), it follows a pattern similar to the throughput results.

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

III. Scenario 3: Deployment density

In the context of the third evaluation scenario, we compare the two mechanisms in a varying environment, in terms of HeNBs' deployment density. The number of the femto cells ranges from 2 (sparse deployment) up to 50 (ultra-dense deployment); the latter resembles a typical 5G scenario as already discussed earlier, and according to [56]. In the figures below (Figure 22-Figure 25) we illustrate the network KPIs that resulted from the third scenario's experimentation. At this point we remind the reader that the *Suitability threshold* of COmpAsS algorithm is set to 0.7, while the *hysteresis* parameter at 0.1.

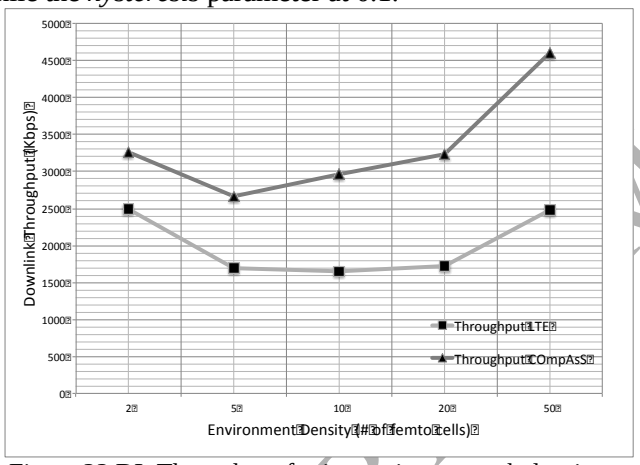

Figure 22 DL Throughput for increasing network density

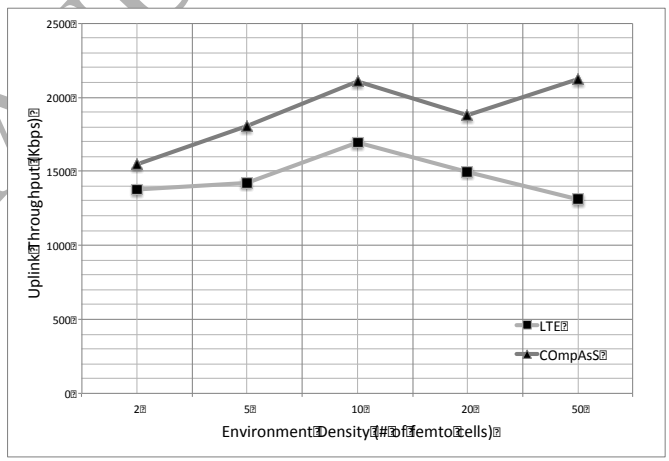

Figure 23 UL Throughput for increasing network density

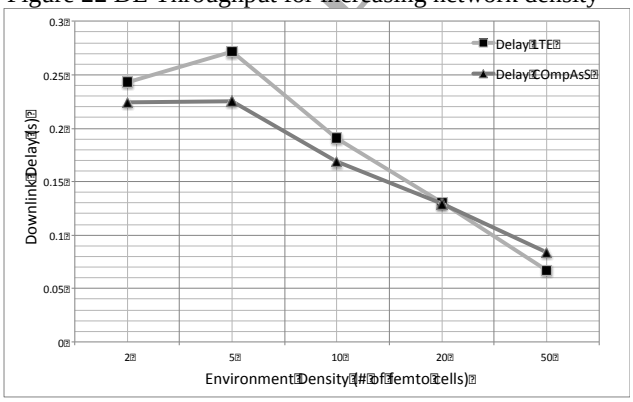

Figure 24 DL Delay for increasing network density

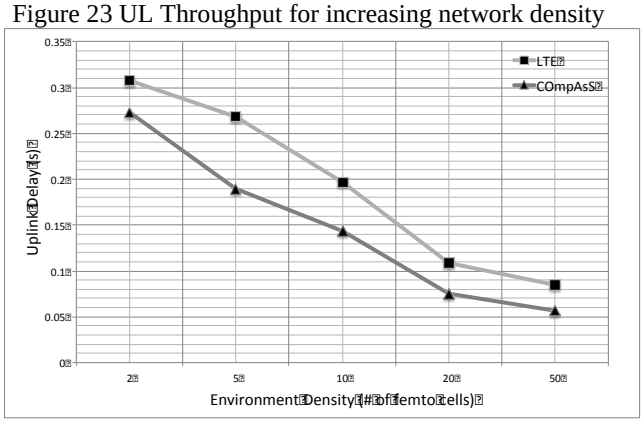

Figure 25 UL Delay for increasing network density

In terms of throughput, COmpAsS achieves higher performance both in the downlink, as well as the uplink. As also illustrated in the respective graph figures, there is no linear relation to the number of available femto cells and the

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

IV. Scenario 4: Network load

In the final round of experiments and in the context of the last scenario, we gradually increase the network load in terms of active bearers (active traffic flows) per UE, aiming at comparing the performance of our scheme in extreme load and interference conditions. In the particular set-up, we deploy 10 femto cells (co-existing with the fixed –throughout all experiment scenarios- number of macro cells). Similarly with the previous scenarios, we provide the downlink and uplink throughput and delay KPIs (Figure 26-Figure 29) for the two mechanisms.

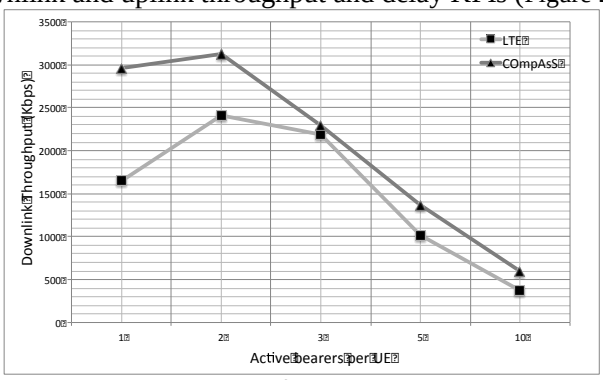

Figure 26 DL throughput for increasing network load

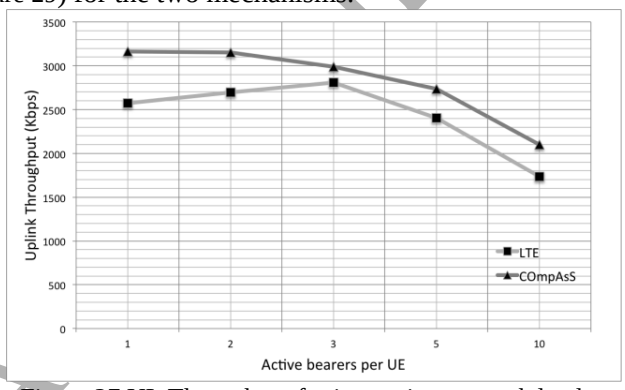

Figure 27 UL Throughput for increasing network load

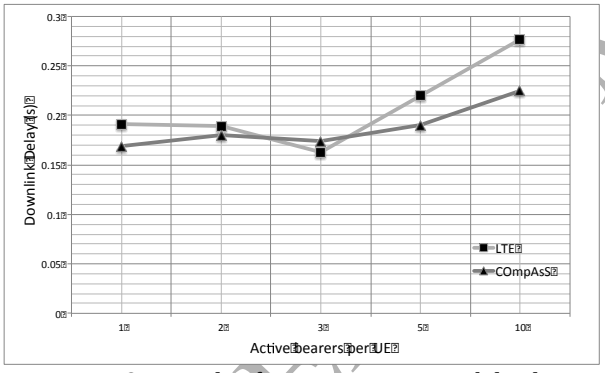

Figure 28 DL Delay for increasing network load

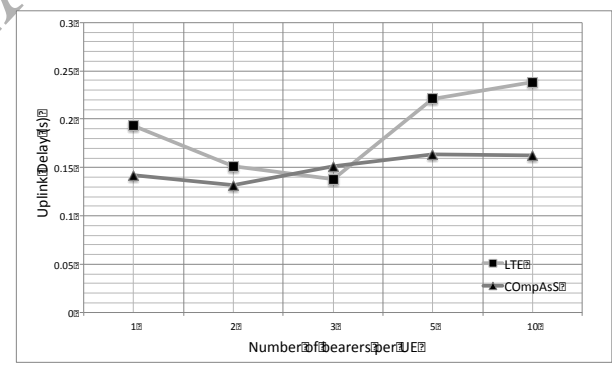

Figure 29 UL Delay for increasing network load

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

V. Conclusion

In the context of this paper we proposed a framework for RAT selection in 5G ultra-dense network environments based on a context-aware, user-oriented scheme, namely COmpAsS. COmpAsS selected the optimal RAT for each

one of the active flows of a UE inside the network. The proposed mechanism collects and subsequently processes context information monitored and aggregated by the UE, taking advantage of latest advancements and 3GPP trends in the LTE-EPC architecture (e.g. ANDSF functionality). This distributed RAT selection framework provides decision-making for RAT selection on a per-traffic flow basis (separately for the uplink and the downlink) and according to each UE's specific service and QoS requirements. Through a comprehensive literature review we demonstrated the added value for COmpAsS. Thereinafter we provided detailed insights with regard to the main components of the proposed scheme, the main algorithmic steps, the fuzzy logic based inference system, as well as the *Suitability* metric for RAT selection. An extensive experimentation series was provided, based on which we assessed the validity and viability of our proposal in a -as much realistic as possible- 5G environment. From a holistic perspective, the proposed UE-oriented framework in the context of the simulations –and via the diverse scenarios that were deployed-, clearly showcased its superiority, in terms of fundamental, QoS-related, network KPIs.

Regarding our future plans, we mainly intend to further elaborate on the orchestration between coexisting context-based frameworks; particularly a central, network-oriented RAT selection and decision-making entity (most probably SDN-enabled, capable of user profiling, mobility prediction, etc.) with a UE-oriented scheme such as the proposed one, leading to a viable, holistic, hybrid solution for the 5G systems.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] S. Sun, M. Kadoch, L. Gong and B. Rong, "Integrating network function virtualization with SDR and SDN for 4G/5G networks," in IEEE Network, vol. 29, no. 3, pp. 54-59, May-June 2015.

[2] X. Ge, S. Tu, G. Mao, C. X. Wang and T. Han, "5G Ultra-Dense Cellular Networks," in IEEE Wireless Communications, vol. 23, no. 1, pp. 72-79, February 2016.

[3] M. Kaneko, H. Yamaura, Y. Kajita, K. Hayashi and H. Sakai, "Fairness-Aware Non-Orthogonal Multi-User Access With Discrete Hierarchical Modulation for 5G Cellular Relay Networks," in IEEE Access, vol. 3, no. , pp. 2922-2938, 2015.

[4] Y. Tao, L. Liu, S. Liu and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," in China Communications, vol. 12, no. 10, pp. 1-15, Oct. 2015.

[5] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[6] Hakiri, Akram and Pascal Berthou. "Leveraging SDN for The 5G Networks: Trends, Prospects and Challenges." Book chapter in "Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture", 2015 John Wiley & Sons, Ltd.

[7] J. Wannstrom, "LTE-Advanced", May 10, 2012, http://www.3gpp.org/IMG/pdf/lte_advanced_v2.pdf [accessed 06/2016].

[8] Nokia, "Smart Wi-Fi Traffic Steering", December 2015

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

[10] 3GPP TS 29.060, V11.5.0, "General Packet Radio Service (GPRS) Tunnelling Protocol (GTP) across the Gn and Gp interface", Release 11, December 2012.

[11] 3GPP Specification Groups, http://www.3gpp.org/specifications-groups

[12] 3GPP TS 23.401, V13.6.0, "GPRS enhancements for E-UTRAN access ", Release 13, March 2016.

[13] 3GPP TS 23.402, V13.5.0, "Architecture enhancements for non-3GPP accesses", Release 13, June 2013.

[14] 3GPP TS 23.852, "Study on S2a Mobility based on GPRS Tunnelling Protocol (GTP) and Wireless Local Area Network (WLAN) access to the Enhanced Packet Core (EPC) network (SaMOG)", September 2013.

[15] 4G Americas White Paper, "Mobile Broadband Evolution Toward 5G: Rel. 12 & Rel.13 and Beyond", June 2015

[16] Apple iOS Connection Manager, https://support.apple.com/en-us/HT202831 [accessed June 2016]

[17] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler and Y. Koucheryavy, "5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells:

Performance Dynamics, Architecture, and Trends," in IEEE Journal on Selected Areas in Communications, vol. 33, no. 6, pp. 1224-1240, June 2015.

[18] M. Ayyash et al., "Coexistence of WiFi and LiFi toward 5G: concepts, opportunities, and challenges," in IEEE Communications Magazine, vol. 54, no. 2, pp. 64-71, February 2016.

[19] Fraunfoher IPMS, Li-Fi, Optical Wireless Communication, http://www.ipms.fraunhofer.de/content/dam/ipms/common/products/WMS/WMS_OWC_2016_web.pdf [accessed 11/2016]

[20] 3GPP TS 29.275, V11.5.0, "Proxy Mobile IPv6 (PMIPv6) based Mobility and Tunnelling protocols", Release 11, December 2012.

[21] 3GPP TS 23.861, V13.0.0, "Network-based IP Flow Mobility", Release 13, June 2015.

[22] 3GPP TS 24.312, V13.2.0, "Access Network Discovery and Selection Function (ANDSF) Management Objects (MO)", Release 13, March 2016.

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

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

[25] "Hotspot 2.0 and ANDSF for Smart Mobile User Connectivity", Altice Labs White Paper, http://www.alticelabs.com/content/WP-Hotspot2-0-and-ANDSF-for-Smart-Mobile-User-Connectivity.pdf, May 2014

[26] IEEE 802.11u-2011, "IEEE Standard for Information Technology-Telecommunications and information exchange between systems-Local and Metropolitan networks-specific requirements" - Amendment 9: Interworking with External Networks.

[27] 3GPP TS 23.203 V13.7.0, "Policy and charging control architecture", Release 13, March 2016.

[28] Wi-Fi Roaming Building on ANDSF and Hotspot 2.0", Alcatel Lucent – BT White Paper.

[29] 3GPP 24.237 V13.4.0, "Mobility between 3GPP Wireless Local Area Network (WLAN) interworking (I-WLAN) and 3GPP systems", March 2016.

[30] 3GPP TS 24.302 V13.5.0, "Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks", Release 13, March 2016.

[31] 3GPP TS 23.261 V13.0.0, "IP Flow Mobility and Seamless Wireless Local Area Network (WLAN) offload", Release 13, March 2016.

[32] The METIS 2020 Project – Laying the foundation of 5G, www.metis2020.com, [accessed July 2016]

[33] 3GPP, TR 23.865, V12.1.0, "Study on Wireless Local Area Network (WLAN) network selection for 3GPP terminals; Stage 2", Release 12, December 2013.

[34] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms", Communications Surveys & Tutorials, IEEE, vol. 16, iss. 1, 1st Quarter 2014, pp. 64-91.

[35] 3GPP TS 36.214, V9.1.0, RSRP and RSRQ Definitions with Receiver Diversity, Release 9, February 2009.

[36] D. Xenakis et. al, "A context-aware vertical handover framework towards energy-efficiency", Vehicular Technology Conference , 2011

[37] K. R. Rao, Z. S. Bojkovic, and B. M. Bakmaz, "Network Selection in Heterogeneous Environment: A Step toward Always Best Connected and Served", Telsiks 2013, October 2013, pp 83-92.

[38] A. Ahmed, L. Merghem Boulahia, and D. Gaıti, "Enabling Vertical Handover Decisions in Heterogeneous Wireless Networks: A State-of-the-Art and a Classification", IEEE Communications Surveys and Tutorials, accepted for publication. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6587998 [retrieved April, 2014].

[39] P. Bellavista, A. Corradi, and C. Gianneli, "A Unifying Perspective on Context-Aware Evaluation and Management of Heterogeneous Wireless Connectivity", IEEE Communications Surveys and Tutorials, vol. 13, No. 3, 3rd quarter 2011, pp. 337 - 357.

[40] X. Yan, Y.A. Sekercioglu, and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks", Computer Networks vol. 54, 2010, pp. 1848-1863.

[41] X. Duan and X. Wang, "Authentication handover and privacy protection in 5G hetnets using software-defined networking," in IEEE Communications Magazine, vol. 53, no. 4, pp. 28-35, April 2015.

[42] H. Song, X. Fang and L. Yan, "Handover Scheme for 5G C/U Plane Split Heterogeneous Network in High-Speed Railway," in IEEE Transactions on Vehicular Technology, vol. 63, no. 9, pp. 4633-4646, Nov. 2014.
[43] J. S. Thainesh, N. Wang and R. Tafazolli, "A scalable architecture for handling control plane failures in heterogeneous networks," in IEEE Communications Magazine, vol. 54, no. 4, pp. 145-151, April 2016.
[44] Y. Kim et al., "Feasibility of Mobile Cellular Communications at Millimeter Wave Frequency," in IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 3, pp. 589-599, April 2016.
[45] H. Zhang, C. Jiang, J. Cheng and V. C. M. Leung, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," in IEEE Wireless Communications, vol. 22, no. 3, pp. 92-99, June 2015.
[46] H. Peng, Y. Xiao, Y. N. Ruyue and Y. Yifei, "Ultra dense network: Challenges, enabling technologies and new trends," in China Communications, vol. 13, no. 2, pp. 30-40, Feb. 2016.
[47] H. A. U. Mustafa, M. A. Imran, M. Z. Shakir, A. Imran and R. Tafazolli, "Separation Framework: An Enabler for Cooperative and D2D Communication for Future 5G Networks," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 419-445, Firstquarter 2016.
[48] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro and K. Zheng, "Toward 5G densenets: architectural advances for effective machine-type communications over femtocells," in IEEE Communications Magazine, vol. 53, no. 1, pp. 134-141, January 2015.
[49] A. Orsino, G. Araniti, A. Molinaro, A. Iera, "Effective RAT Selection Approach for 5G Dense Wireless Networks", Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st.
[50] C.Y. Luiu, F.Y. Liu, A. Castiglionne, F. Palmieri, "Heterogeneous Network Handover Using 3GPP ANDSF", 2015 IEEE 29th International Conference on Advanced Information Networking and Applications, March 2015, pp. 171-175
[51] D. Xenakis, N. Passas, L. Merakos, C. Verikoukis, "ANDSF-Assisted vertical handover decisions in the IEEE 802.11/LTE-Advanced network", Elsevier Computer Networks Journal vol. 106, 2016, pp. 91-108.
[52] L. Xia, L Jiang, and C. He, "A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method", IEEE ICC, June 2007, pp. 5665 - 5670.
[53] B. Ma and X. Liao, "Vertical Handoff Algorithm Based on Type-2 Fuzzy Logic in Heterogeneous Networks", Journal of Software, vol. 8, No 11, November 2013, pp. 2936-2942.
[54] 3GPP 23.890, V 12.0.0, "Optimized offloading to Wireless Local Area Network (WLAN) in 3GPP Radio Access Technology (RAT) mobility", September 2013.
[55] 3GPP 36.304, V.13.1.0, "User Equipment (UE) procedures in idle mode", Section 5.2.4.3 "Mobility states of a UE", Release 13, March 2016.
[56] METIS EU Project, https://www.metis2020.com
[57] 3GPP TR 36.931 v13.0.0, "Radio Frequency (RF) requirements for LTE Pico Node B", Release 13, January 2016
[58] 3GPP TS 36.921 v13.0.0, "FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis", Release 13, January 2016
[59] NS3 online documentation, https://www.nsnam.org/docs/models/html/lte-design.html#fig-lte-legacy-handover-algorithm, [accessed July 2016]
[60] Traffic model related information, http://www.statista.com/statistics/185868/average-mobile-wireless-call-length-in-the-united-states-since-june-1993/
[61] Traffic model related information, http://en.wikipedia.org/wiki/Adaptive_Multi-Rate_Wideband
[62] Traffic model related information, http://networks.nokia.com/system/files/document/volte_white_paper_final.pdf
[63] Traffic model related information, http://www.swisscom.ch/dam/swisscom/en/res/mobile/mobile-network/netztest-connect-en-2014.pdfhttp://www.swisscom.ch/dam/swisscom/en/res/mobile/mobile-network/netztest-connect-en-2014.pdf
[64] Traffic model related information, http://www.theglobeandmail.com/technology/tech-news/how-much-bandwidth-does-streaming-use/article7365916
[65] Traffic model related information, http://www.sysomos.com/reports/youtube
[66] Traffic model related information, http://www.broadbandgenie.co.uk/mobilebroadband/help/mobile-broadband-usage-guide-what-can-you-get-for-your-gigabyte
[67] D. Drajic, S. Krco, I. Tomic et al., "Traffic generation application for simulating online games and M2M applications via wireless networks", 2012 9th Annual Conference on Wireless On-demand Network Systems and Services (WONS), IEEE, pp. 167-174

Mr. Sokratis Barmpounakis obtained his five-year Engineering Diploma from National Technical University of Athens (NTUA), in the Department of Electrical and Computer Engineering. After completing his studies, he moved to Geneva, Switzerland. At University of Geneva, he worked for 2 years as a Researcher/Developer in various projects, both Swiss and EU. Since March 2013, he is a PhD candidate in the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens (UoA). His main fields of interest are 5G, Mobile Networks, Context-Aware Mobility and Resource Management and Collaborative Data Sharing.

Prof. Alexandros Kaloxylos received the B.Sc. degree in Computer Science from the University of Crete, Greece, in 1993, the M.Phil. degree in Computing and Electrical Engineering from the Heriot-Watt University, Scotland, in 1994 and the Ph.D. degree in Informatics and Telecommunications from the University of Athens in 1999. He has participated in numerous projects realized in the context of EU programs as well as National Initiatives. He has currently published over 80 papers in international journals and conferences. He is a senior member of IEEE and a member of the editorial board of the IEEE Communication's Society Surveys and Tutorials Electronic Journal.

Dr. Panagiotis Spapis received his diploma in Electrical and Computer Engineering from the University of Patras, Greece, 2008. In 2015, he received his PhD from the Department of Informatics & Telecommunications of UoA. His research focused on learning-enhanced situation awareness and perception in future wireless networks. From 2008 until 2015 he served as researcher at the Software Centric & Autonomic Networking (SCAN) group of National and Kapodistrian University of Athens, where he has participated in numerous research projects funded by the EU. His

main research interests lie in the area of context aware, learning enhanced mechanisms for 5G networks. Parts of his work have already been published in several conferences and journals.

Prof. Nancy Alonistioti has a B.Sc. degree and a PhD degree in Informatics and Telecommunications (Dept. of Informatics and Telecommunications, University of Athens). She has working experience as senior researcher and project manager in the Dept. of Informatics and Telecommunications at University of Athens. She has participated in several national and European projects, which had a focus on reconfigurable mobile systems, cognitive mobile networks and FI. She is co-editor and author in “Software defined radio, Architectures, Systems and Functions”, published by John Wiley in May 2003. She is TPC member in many conferences in the area of mobile communications and mobile applications for systems and networks beyond 3G.