Softwarization and virtualization in 5G mobile networks: benefits, trends and challenges

Massimo Condoluci, *Member, IEEE,* Toktam Mahmoodi, *Senior Member, IEEE,*

Abstract

The promise behind the effective deployment of 5G networks is an architecture able to provide flexibility, reconfigurability and programmability in order to support, with fine granularity, a wide and heterogeneous set of 5G use cases. This dictates a radical change in the design of mobile systems which, being usually based on the use of static deployment of vendor equipment characterized by monolithic functionality deployed at specific network locations, fail in providing the above mentioned features. By decoupling network functionalities from the underlying hardware, softwarization and virtualization are two disruptive paradigms considered to be at the basis of the design process of 5G networks. This paper analyses and summarizes the role of these two paradigms in enhancing the network architecture and functionalities of mobile systems. With this aim, we analyze several 5G application scenarios in order to derive and classify the requirements to be taken into account in the design process of 5G network. We provide an overview on the recent advances by standardization bodies in considering the role of softwarization and virtualization in the next-to-come mobile systems. We also survey the proposals in literature by underlining the recent proposals exploiting softwarization and virtualization for the network design and functionality implementation of 5G networks. Finally, we conclude the paper by suggesting a set of research challenges to be investigated.

Index Terms

5G, mobile core, softwarization, virtualization, SDN, NFV.

I. Introduction

The evolution of mobile networks from 2G to 4G has been mainly driven by the supporting applications, whose requirements defined the features of the network in terms of procedures (e.g.,

Massimo Condoluci is with Ericsson Research, Sweden, e-mail: massimo.condoluci@ericsson.com & Toktam Mahmoodi, is with the Centre for Telecommunications Research, Department of Informatics, King's College London, UK, e-mail: toktam.mahmoodi@kcl.ac.uk.

authentication, signaling, connection establishment) and functionalities (e.g., mobility management, anchoring, data forwarding, path computation). Once the application-related features have been defined, mobile networks have been deployed by exploiting the network technologies, which were well-consolidated [1]. As each generation of mobile networks has a design to deliver specific services, the introduction of novel applications to satisfy customers' demands requires the re-design or the introduction of novel functionalities in the network in case the requirements of such applications differ substantially from those of the applications already supported [2]. From this point of view, the well-consolidated use from 2G to 4G networks of hardware implementing specific network functionalities has the following drawbacks [3], [4], [5]: *(i)* updating the already deployed network functions requires the introduction of novel hardware equipment; *(ii)* supporting novel applications could require totally disruptive network changes, and this dictates for a novel network design and thus the standardization of a novel generation of mobile networks. In addition to the mentioned limitations, high capital expenditure (CAPEX) in deploying new network architectures and operating expenditure (OPEX) when upgrading network functionalities have been extra barriers that consequently decreases the revenue of mobile operators [6].

Above mentioned aspects are exacerbated when considering 5G networks, which have to effectively support the Internet of Things (IoT) and all its related services [7]. Indeed, in addition to the traditional human-type applications (such as video streaming, web browsing, VoIP, video conferences, etc.), 5G networks are expected to support a wide range of applications coming from the autonomous communication among sensors and actuators [8], [9] as well as the interaction between humans and machines [10], [11]: smart cities, intelligent mobility, tactile Internet, industry automation, remote surgery are only few examples of the application environments envisioned for 5G networks. It thus becomes clear that, in addition to the simultaneous support of human-type communications (HTC) and machine-type communications (MTC) with profound differences in terms of packet size, traffic pattern, energy consumption, etc. [12], [13], the next-to-come generation of mobile networks has to face with applications with heterogeneous characteristics and Quality of Service (QoS) requirements. For example, intelligent mobility involves vehicles moving at different velocity speeds and thus requires ad-hoc solutions to properly handle mobility management [14]; on the contrary, industry automation mainly deals with static devices and thus mobility management features can be relaxed but ultra-low latency needs to be guaranteed to avoid chain instabilities [15]. Flexibility is thus one of the main drivers

in the design of 5G networks in order to support heterogeneous services: the architecture and the functionalities of 5G networks need to be flexible in order to simultaneously accommodate the heterogeneous requirements of 5G applications.

The provisioning of flexibility requires a drastic change in the technologies, which is even more highlighted when linked with lowering CAPEX and OPEX. In radical contrast with hardware-based deployments of previous generations, the design of 5G networks takes advantage of the exploitation of software-based technologies which are gaining importance in these recent years. Indeed, two novel paradigms, i.e., *softwarization* and *virtualization* [16], [17], are targeted as two disruptive technologies to be adopted in 5G networks to allow flexibility and fast re-configuration of the network, based on the delivered services. The joint use of softwarization and virtualization allows network functions to run in software instead of hardware and decouples network functions from the underlying hardware platform: this allows functionalities to be upgraded and installed in any location of the network in order to meet the requirements of the services to be delivered.

As softwarization and virtualization are two technologies that have been recently considered in mobile networks, the aim of this paper is to provide an overall overview of the needs of 5G applications in order to highlight the role that such technologies have in the still ongoing design of 5G networks. The main features of softwarization and virtualization are discussed in more details in the remainder of this Section, in order to provide a better understanding of these paradigms to the reader. Finally, Sec. I-B lists the detailed contributions of this paper and its organization.

A. *Softwarization and virtualization*

Softwarization refers to the paradigm where a given functionality runs in software instead of hardware. This approach guarantees high degree of flexibility and reconfigurability as functionalities can be enhanced (or novel functions introduced) by updating the software. As softwarization breaks the relationship between functions and the underlying hardware, it has obvious benefits in terms of time to deploy, CAPEX and OPEX reduction when updating or introducing new functions. *Virtualization* enhances the software/hardware splitting of the softwarization paradigm by creating abstracted (i.e., virtual) instances of hardware platforms, operating systems, storage devices, and computer network resources. This means that, with virtualization, software runs in commercial off-the-shelf (COTS) equipment (e.g., standard server) by exploiting a virtual machine (VM) instead of a dedicated hardware. As VMs can be moved across different hardware

platforms, virtualization introduces flexibility by means of flexible VM placement and migration. Other benefits are in terms of reduced CAPEX and OPEX, as the introduction of new services requires the introduction of new VMs without requiring any effort from a hardware point of view. Virtualization is at the basis of cloud computing technologies, offering on-demand access to different applications and services by sharing pool of configurable computing resources (e.g., networks, servers, storage). A cloud environment can offers services for all types of managed resources (from computation/storage capabilities to software), this is the everything as a service (XaaS) model, which can be particularized into the following models: Infrastructure as a Service (IaaS), i.e., the provisioning of processing, storage, networks, and other fundamental computing resources to users which run their own applications and have control over operating systems, storage, and deployed applications; Platform as a Service (PaaS), where users can deploy their own applications but without any control of the underlying cloud infrastructure; Software as a Service (SaaS), i.e., the provisioning of applications running on a cloud infrastructure.

From a network point of view, Software Defined Networking (SDN) is considered as the main realization of the softwarization concept [18]. SDN is composed of the following components:

- *SDN applications.* These are programs (i.e., software) which run network functionalities (e.g., routing, link/path configuration).
- *SDN controller.* It is a logically centralized entity which provides the SDN applications with an abstract view of the network (also with additional information in terms of statistics and events) and translates the inputs received from the SDN applications down to the physical network (i.e., switches).
- *SDN datapath.* It is composed by several network devices, which are configured and instructed by the SDN controller to perform actions (such as packet forwarding).

The interface between the SDN controller and applications is referred to as the northbound interface, while the SDN controller communicates with the underlying network devices by means of the southbound interface. The northbound interface is typically implemented by means of a REpresentational State Transfer (REST) interface using the Yet Another Next Generation (YANG) model and XML or JavaScript Object Notation (JSON) languages to communicate with the SDN controller [19]. From a southbound interface point of view, the Open Networking Foundation (ONF)[1] was founded in 2011 by several companies aimed at improving networking

[1]https://www.opennetworking.org/about/onf-overview

through SDN by means of developing a common southbound interface, namely the OpenFlow protocol, which is currently considered as the de-facto standard southbound interface for SDN [20]. SDN introduces the possibility of decoupling network control from forwarding functions by thus allowing flexibility and reconfigurability of the physical network [21]. With a proper configuration of switches' queues and meters, SDN is able to provide isolation among different traffic types or paths, and thus to effectively slice network resources. For more details on SDN, please refer to [16], [21], [22], [23].

From a network point of view, the exploitation of the virtualization paradigm brings to the concept of Network Function Virtualization (NFV), proposed in 2012 by the European Telecommunications Standards Institute (ETSI) [24]. With NFV, network node functions (such as firewall, switches, etc.) are virtualized and thus totally decoupled from the underlying hardware running such functions. The three main components of NFV are:

- Virtualized Network Functions (VNFs). A VNF is the software implementation of network functions which can be composed of one or more VMs. A set of VNFs executed in a specific order is referred to as Service Function Chain (SFC).
- Network function virtualization infrastructure (NFVI). This includes all the hardware and software components that build the environment where VNFs are deployed.
- Network functions virtualization management and orchestration (NFV-MANO) Architectural Framework. This is usually composed of a: Virtual Infrastructure Manager (VIM), controlling physical and virtual infrastructures; VNF manager, handling VNFs and SFCs and their placement; orchestrator, in charge of managing the VIM and the VNF manager [25].

For more details on NFV, please refer to [26], [27], [28]

Softwarization and virtualization can be thus seen as two complementary technologies to provide flexibility in 5G networks, where cloud-based environments can be used to run and move on-demand VNFs while SDN can dynamically change the network topology according to the load and service requirements [29]. The benefits of softwarization and virtualization in 5G networks are discussed in details in Sections III-B and III-C, while Sec. IV focuses on how SDN and NFV technologies are currently considered by the current literature as key enablers of 5G networks.

B. Contributions and organization of the paper

The remainder of this paper is organized as follows. Sec. II summarizes the evolution of mobile networks from 2G to 4G. This Section contributes to underline the close relationship between the requirements of services expected to be provided by mobile networks and the design drivers at the basis of their architecture, mainly inherited from the well consolidated solutions adopted in fixed networks to handle such services. We further highlight the relationship of each generation of mobile networks with the previous one, in order to provide a detailed picture of current 4G networks by describing the reasons behind the implemented network functionalities and the deployed architecture entities. Finally, we provide a summary of the limitations of current 4G networks.

In Sec. III we provide an overview of the use cases of 5G networks and we classify their related requirements. We further present the architectural enhancements considered as key features of 5G network architecture.

The contribution of Sec. IV is to provide a comprehensive overview of the state of the art about softwarization and virtualization in 5G networks. We survey the different works in literature that deal with softwarization and virtualization from different angles, ranging from network architecture to function definition and placement.

Sec. V highlights the research challenges that still need to be investigated by considering the research efforts discussed in Sec. IV.

Finally, Sec. VI gives the conclusive remarks.

II. Evolution of the Mobile Network Architecture

In the last decades, mobile network architecture evolved from 2G to 4G to support always more high-demanding services. In order to clearly analyse the design features at the basis of the current 4G networks, in this Section we provide a description of the mobile architecture evolution, depicted in Fig. 1. The aim is to underline the role that applications to be supported and available network technologies had in defining the design choices of each generation of mobile networks. Finally, we present an overview of the studies in literature that discuss the limitations of 4G architecture.

A. From phone calls in 2G to mobile data in 3G

The shift from 2G to 3G was mainly to introduce mobile Internet and additional capabilities and features to the 2G network. In details, the Global Systems for Mobile Communications

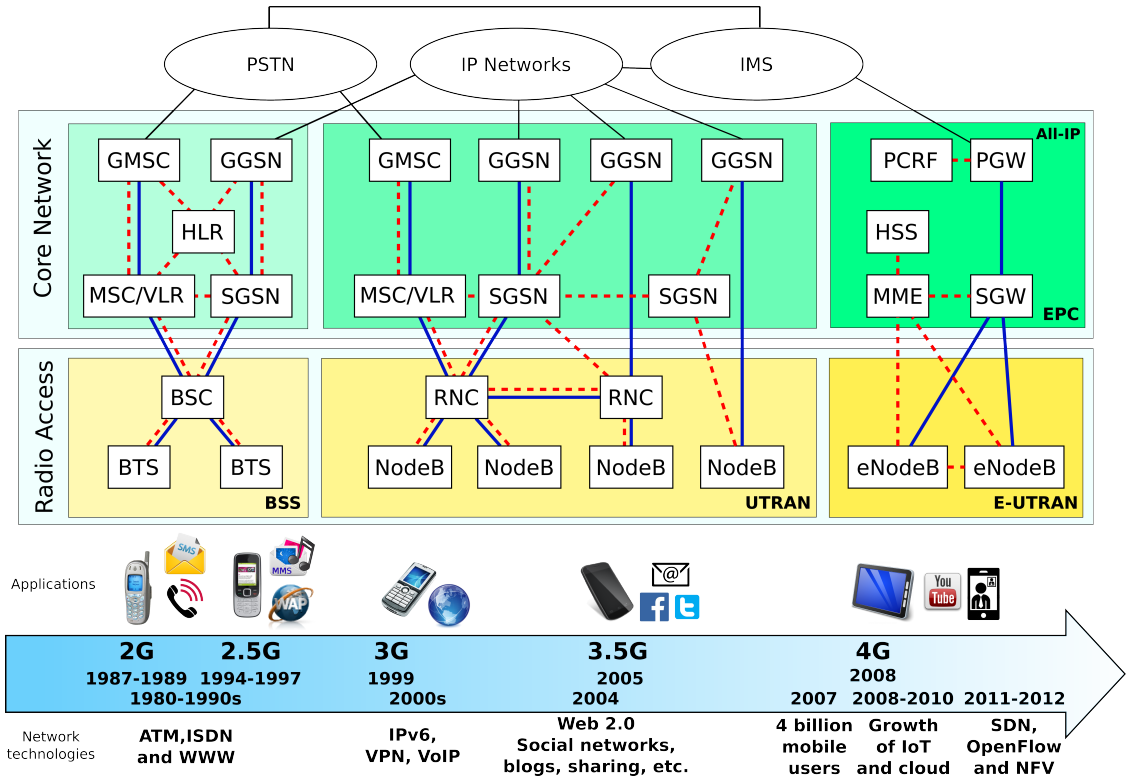

Fig. 1. The evolution of mobile network architecture from 2G to 4G. This drawing highlights the applications supported by each generation of mobile networks as well as the network technologies available during the standardization process of mobile networks. The blue lines represent the data plane traffic, while red ones refer to the control traffic.

(GSM), a.k.a. 2G, was first developed in the 1980s with the main purpose of providing voice calls to mobile users [30], [31]. GSM adopted the same well-consolidated solutions to manage voice calls over fixed telephone network, i.e., *circuit-switched* digital communications, for an easy compatibility between fixed and mobile networks [32]. From a Radio Access Network (RAN) point of view, a.k.a. Base Station Subsystem (BSS) in the GSM terminology, 2G networks rely on the Base Station Controller (BSC), in charge of handling the management of radio resources, handover, power management and signalling for a set of base station (BS)s, a.k.a. Base Transreceiver Subsystem (BTS). The Core Network (CN), namely the Network Subsystem (NSS), is composed of the Mobile Switching Centre (MSC) that is responsible of controlling a set of BSCs [33] and configuring user traffic by means of the Home Location Register (HLR), a database storing subscription information of mobile users. Finally, the edge node of the 2G core network is the Gateway Mobile Switching Centre (GMSC), inter-connecting the GSM network to external Public Switched Telephone Network (PSTN) and Integrated Services Digital Network

(ISDN) networks[2]. It is worth noticing that all network entities handle both user traffic, a.k.a. *user plane* (U-plane), and control traffic, a.k.a. *control plane* (C-plane). In the 1990s, the General Packet Radio Service (GPRS), a.k.a. 2.5G, in order to provide data connectivity over mobile networks through Wireless Application Protocol (WAP) [34], [35]. From a network point of view, the breakpoint of GPRS compared to GSM is mainly in the CN and is the exploitation of *packet-switched* communications carried through Asynchronous Trasnfer Mode (ATM) protocol, a consolidated solution to handle bursty-based data traffic in fixed networks [36]. From an architecture point of view, as it can be seen from Fig. 1, GPRS introduced the Serving GPRS Support Node (SGSN) that provides the same functionalities as the MSC for packet-switched traffic. In addition, GPRS introduced the Gateway GPRS Support Node (GGSN) to provide connectivity with external packet switched networks [37][3].

As the market request in the 1900s was mainly focusing on having higher data rates for web browsing through HTTP, the standardization of 3G mobile networks mainly focused on overcoming the poor channel capacity of 2.5G[38]. The main enhancement of 3G, widely referred to as Universal Mobile Telecommunications Service (UMTS) as defined 3rd Genration Partnership Project (3GPP) Rel. 99, is thus the exploitation of Wideband Code Division Multiple Access (WCDMA) on the radio channel for a better utilization of radio resources [39]. The access network, namely the Universal Terrestrial Radio Access Network (UTRAN), does not show meaningful changes compared to 2G architecture as highlighted in Fig. 1[4], except for the introduction of enhanced functionalities at the eNodeB (e.g., rate adaptation, spreading/de-spreading, closed-loop power control) and the Radio Network Controller (RNC) (load and congestion control, open-loop power control, etc.) From a CN point of view, 3G networks are still conceived as an extension of 2G/2.5G networks as underlined by the presence of both circuit- and packet-switched domains [40].

The network architecture started to change with High Speed Packet Access (HSPA), i.e., 3GPP Rel. 5 and 6 (a.k.a. 3.5G). As the aim was still the support of higher data rates, HSPA started

[2]It is interesting to underline the fact that communications between the MSC and the GMSC are handled by grouping multiple GSM phone calls flows (with a data rate of about 13kbps) into one single 64kbps flow (the legacy data rate of voice calls over fixed networks): this aspects underlines that GSM has been designed to mainly cope with only one application, i.e., voice calls.

[3]It is worth noticing that architecture convergence was not considered when introducing 2.5G, i.e., circuit-switched CN handling voice calls and packet-switched CN handling data traffic are simultaneously present.

[4]For the sake of clarity, the HLR is not depicted for 3G and 3.G networks.

to improve the RAN segment (for instance by moving functionalities such as radio resource management from the RNC to the NodeB [41], [42]) as the available CN was still sufficient to handle the amount of data traffic of mobile users [42]. In the 2000s, the availability of always increasing data rates brought to an exponentially increase in the number of mobile users; this pushed the need for further changes of the 3.5 CN. With this aim, in the Rel. 7, 3GPP introduced the so-called direct tunnel solution allowing the U-plane to bypass the SGSN, which is thus only involved in handling the C-plane [42].

B. The 4G architecture

The shift from 3G to 4G was mainly motivated by 2000s need of high-speed data connectivity for mobile web browsing as well as to support real-time Internet traffic such as voice calls, VoIP, video streaming, etc. [43].

The industry has converged on a new standard for mobile communications, defined in 3GPP Rel. 8, to effectively support human-oriented Internet applications with strict QoS needs (such as video calls, video streaming, etc.). 3GPP Rel. 8 defines the Evolved-UTRAN (E-UTRAN), a.k.a. Long Term Evolution (LTE), and the Evolved Packet Core (EPC) of the so-called 4G network[5]. In addition to the use of a novel radio interface based on Orthogonal Frequency Division Multiplexing (OFDM), to boost data rates and to improve the management of radio resources [45], [46], 4G networks introduced a radical change from an architectural point of view, i.e., one single CN handling both voice and data traffic types. This was possible thanks to the recent advances in handling QoS over IP-based networks and motivated the use of IP in the mobile core instead of ATM-based solutions for data traffic and circuit-switched networks for voice calls. So doing, 4G networks are customizable from a traffic point of view, by exploiting IP QoS features to manage heterogeneous human-type traffic types with different QoS features in terms of packet losses, maximum tolerated delay and priority. 3GPP Rel. 8 standardized the first *flat all-IP* mobile architecture where IP is exploited as a network layer to inter-connect RAN and CN entities regardless the underlying layers and technologies as well as the traffic type.

The 4G architecture enhances the C/U-plane splitting started with 3.5G [47]. For both planes, the anchor point is represented by the Evolved NodeB (eNodeB), the only entity in the E-UTRAN.

[5]For the sake of completeness, LTE Rel. 8 is considered to belong to 3.9G, as it does not fulfil the requirements defined in International Mobile Telecommunications-Advanced (IMT-A) over the radio interface. The real 4G radio interface is LTE-Advanced (LTE-A), although from an commercial point of view LTE is branded as 4G technology [44].

In the EPC, the Serving Gateway (SGW) and the Packet Data Network Gateway (PGW) mainly work in the U-plane. The SGW is the termination point of the packet data interface towards the E-UTRAN and acts as a local mobility anchor in case of handover. The PGW is the termination point of the packet data interface towards external networks. Furthermore, the PGW also supports policy enforcement features, packet filtering (e.g., deep packet inspection) and evolved charging support; to this end, the PGW exploits the Policy and Charging Rules Function (PCRF). In the C-plane, the Mobility Management Entity (MME) supports: *(i)* security procedures such as end-user authentication, initiation and negotiation of ciphering and integrity protection algorithms; *(ii)* signalling procedures, used to set up packet data context; *(iii)* location management, i.e., tracking area update process. The MME exploits the Home Subscriver Server (HSS) to request subscription information of mobile users.

Although the 4G architecture yields to easier management and scalability compared to previous generations of mobile core, this architecture presents different limitations in terms of flexibility [48]. The remainder of this Section will focus on the limitations of the 4G architecture.

C. Limitations of 4G mobile core

From an architectural point of view, the design drivers of 4G networks (e.g., use of static deployment of vendor equipment and use of monolithic functionality at specific network locations inherited from previous mobile generations) introduce different limitations to successfully extend the set of applications supported by 4G networks. All these limitations, summarized in Table I, can be classified as follows.

1) Inflexibility: Although current 4G networks are more customizable compared to 3G, the presence of proprietary black boxes, huge variety of expensive and proprietary equipment and inflexible hard-state signalling protocols limits the programmability of the network and thus its flexibility [3], [4]. EPC elements are controlled through standardized interfaces and cannot be controlled by open interfaces or through Application Programming Interface (API)s [49]. As a consequence, the introduction of new services and functionalities is not possible without heavy integration within operators' network: this led to high deployment (integration of new services or the change of network topology and architecture would require the replacement or redistribution of hardware equipment running the network functionalities) and operational costs (mobile operators have to replace existing equipment even if it is still sufficient for most purposes) which are not any more affordable considering the current mobile communication [50].

2) Complexity: In terms of functionalities, the 4G architecture shows a clear C/U-plane separation in the E-UTRAN, but C-plane functions are frequently co-located with U-plane equipment in the backhaul/core network as for instance it happens with SGW and the PGW. Although the SGW is the U-plane anchor-node for the eNodeB and the PGW is considered as the edge-node to allow data interconnection with external networks, these nodes carry U-plane traffic and manage C-plane signalling messages.

Additional issues in terms of complexity need to be considered as they could involve delays for C/U-planes [51], [52]. By considering the PGW, this entity is in charge of establishing and managing the bearers that allow the User Equipment (UE) to be connected to the network and this means that all the traffic should pass through the PGW, even if the communication is between UEs attached to the same cell [53], [52]. As a consequence, strategies like caching of popular content in the mobile network find limited benefits (only in terms of performance improvements for end-users) in the current 4G deployment because all flows have to pass through the PGW [49].

Another aspect that needs to be taken into consideration is the *tunneling* between network elements. Although being an all-IP network, IP packets in the EPC are always tunnelled by using the GPRS Tunneling Protocol (GTP) [47]. This solution is inherited by previous generations networks as it guarantees to inter-connect different type of networks [17]: convergence as "offering a converged service view" [4] is a key design driver for 3G/4G networks, and the use of tunneling was useful in this direction. Nevertheless, next-to-come architecture is expected to deal with *infrastructure convergence* [4], and this may require a re-consideration in the exploitation of tunneling. The control and the management of tunnels introduce complexity in the network and reduce its scalability [54], and this is thus translated into additional costs for mobile operators [17].

Finally, it is worth to remark that the actual structure of EPC and E-UTRAN may become an obstacle to future evolutions of the network when considering the possibility to introduce new technologies and services [49]. A major example is the applicability of the SDN paradigm and, in this context, scalability is the most limiting aspect in turning into reality the vision of SDN in EPC. Indeed, a typical PGW can support a few million users[6] and, for each user there is the

[6]Cisco packet gateway. http://goo.gl/XM6dCt.

need to handle unique charging, QoS and forwarding rules[7]: this means having tens of millions of rules to be applied at line rate. Although each switch will manage only a portion of these rules, the state-of-art programmable switching platforms can only support tens of thousands of forwarding rules[8] and this number may be too small according to the capacity and deployment of current and future cells.

Another scalability issue is in terms of signalling traffic. According to Nokia [55], signalling is growing 50% faster than data traffic in LTE networks [56]. As the EPC has a centralized control traffic management, the MME can reach a traffic volume in the C-plane of about 290,000 messages per second in networks with millions of users [55]. As a consequence, the current 4G design introduces scalability issues when considering the possible adoption of SDN to the mobile core [57] and, thus, this dictates for novel design criteria for 5G mobile core.

3) Centralized U/C-planes: An additional limitation for 4G networks is the exploitation of centralized U/C-planes. The U-plane centralization has the benefits of seamless L3 mobility and allows for easy implementation of backward compatibility with older 3GPP technologies. Nevertheless, centralized U-plane also means that ad-hoc solutions are needed to determine the geographical location of UEs with reasonable fidelity to provide geographic customization of services [47], [58].

Additional inefficiencies are associated to the management of 4G U-plane in terms of overhead. The initial attachment procedures lead to a U-plane establishment (i.e., the creation of GTP tunnels between eNodeB, SGW and PGW) even when there is no data traffic to be sent [55]. Another aspect to be remarked is that U-plane parameters are unaware of the session type (an example is the fact that the UE inactivity timer is a static value configured at the eNodeB instead of being dynamically adapted to the session profiles) [47]. A drawback is thus that the UE inactivity timer could expire even if the UE does not have data to send. According to above considered issues, maintaining the U-plane in 4G networks may involve high signalling overhead and this limits the efficiency of the network [62]. This also limits the support of services, such as MTC, with unique features in terms of energy consumption and signalling overhead [63], [64].

[7]Ericsson evolved packet gateway. http://goo.gl/tEjcWP

[8]Intel Ethernet switch fm6000 series. http://goo.gl/FwHF7D

TABLE I

SUMMARY OF LIMITATIONS OF 4G MOBILE NETWORKS

Limitations	Motivations	Drawbacks
Inflexibility	• Static deployment of vendor equipment [3] • Use of proprietary black boxes [3] • Use of inflexible hard-state signalling protocols [49] • Use of monolithic functionalities [3] • Reconfiguration/updates available only with heavy integration within network operators [50] • Updates require replacements of existing equipment even if it is still sufficient for most purposes [50]	• High CAPEX to extend the network (i.e., increasing coverage) [50], [59], [60], [61] • High CAPEX to introduce novel services [50], [59], [60], [61] • High OPEX for network management [50], [59], [60], [61]
Complexity	• Network entities involved in both U/C-planes [51], [52] • Use of GTP tunnels [54], [17] • Offering a converged service view [4] • Centralized U-plane (issues at the SGW and PGW) [53], [49] • Centralized C-plane (issues at the MME) [55] • High number of information related to UEs (charging, QoS, etc.) [47]	• Delay susceptible of the overload at the MME, SGW and PGW [53], [49], [55] • Reduction of the overall capacity (in terms of resources to be used for data traffic) of the network [62] • High CAPEX/OPEX to scale the network [17]
Centralised U/C-planes	• U-plane unaware of session profiles and characteristics [47] • C-plane signalling is growing 50% faster than data traffic [55]	• U-plane connectivity may involve a waste of network resources even if UEs do not have data to send [62] • Reduction of the overall capacity (in terms of resources to be used for data traffic) of the network [62] • Ad-hoc solutions required to get fine-grain UE's information (exact location, etc.) [47], [58] • High delay for control procedures [17] • Energy consumption due to control signalling cannot be considered negligible for some services (e.g., MTC) [63]

III. Roadmap of Softwarization and Virtualization in 5G Environments

While the move to 4G was mainly triggered by the need higher capacity and faster connectivity, the move to 5G is motivated by the diversity of use cases that are expected to be supported. This Section provides a description of expected 5G use cases in order to derive the requirements to be considered in the design of 5G architecture and also describes the role of softwarization and virtualization in satisfying these requirements.

A. Application Environments

Following the boost of IoT, 5G networks are expected to natively support MTC, i.e., the traffic generated by devices without the human intervention [65].

One application scenario fpr 5G and IoT is the *smart sustainable city*. Zanella et al. [66] discusses urban IoT technologies that are close to standardization, and agree that most of smart city services are based on a centralized architecture where data is delivered to a control centre in charge of subsequently processing and storing the received traffic. Within smart cities, *intelligent mobility* is one of the challenging scenarios where autonomous or assisted driving cars need to continuously monitor the situation outside and inside the car and exchange information between the different participants of the transport network, i.e., vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications. Other services in a smart city involve management of traffic congestion, pollution monitoring, parking, etc. Thus, the main task of 5G is to integrate the management of these very diverse services and devices in an efficient manner, by taking into account the diverse nature of the devices (e.g., cars moving at different mobility speeds and fixed traffic road sensors). Adaptability and re-configurability of SDN/NFV is beneficial smart cities, where quick re-configurations of network parameters according to traffic state would improve the capabilities in supporting and optimizing smart cities services [67]. For example emergency services could be deployed on commercial network while re-configurability of SDN/NFV will allow for ensuring the guarantees needed for such emergency services. To enable such re-configurability, data from smart city services can be exploited, such that smart city network and services could reside in *symbiotic* manner. The concept of such co-existence, i.e. *Symbiocity* as well as optimization and self-organization of the network using smart city data have been explored in [68] and [69].

Remote control is an application that can be seen as an evolution of the IoT paradigm to remotely control the objects connected to the network [10]. Remote control is an application that

can play a role in many industries: remote diagnosis and intervention in healthcare industries [70] or remote control of heavy machineries for factory automation [71]. Among those, the *healthcare* industry is showing a strong trend in exploiting remote control to enable services such as decentralization of hospitals, where medical care can be provided at home or while moving, or remote surgery [72], [73]. The applications dealing with remote control often fall in the category of mission critical MTC, with stringent network requirements, i.e., ultra-low delay and ultra-high reliability). For remote control applications, SDN and NFV play a central role in guaranteeing low latency. NFV is able to place or move functionalities needed for remote control closer to the edge of the network, thus cutting the delay in reaching remote servers hosting control applications. SDN is able to guarantee isolation of network resources, thus avoiding that remote control applications will be affected by the presence of other services.

B. Capabilities added to the 5G networks

In this Section, we analyse the capabilities to be considered in the design of 5G networks according to 3GPP [74] to handle the application environments discussed in Sec. III-A; we further discuss the role of softwarization and virtualization in allowing such changes. A high-level 5G architecture is depicted in Fig. 2. The main capabilities to be added into the 5G mobile core are mapped to the general requirements for the business applications discussed above in Tab. II. More details are given in the following of this Section.

1) Flexibility and programmability: From the above use case review, it is clear that *flexibility* is one of the key features to be introduced in 5G networks in order also to support next-to-come use cases developed once 5G will be deployed. It is thus necessary to evolve into a communications network that is sufficiently broad and it is as well able to cope with new applications or other verticals not being considered yet. This is considered to be one of the main features of 5G, i.e., *programmability*, to allow network functions to be extended or re-designed in order to support new use cases or enhanced functionalities.

The role of softwarization and virtualization is of primary importance in the introduction of flexibility and programmability in the 5G architecture and functionalities of various building blocks. In other words, softwarization introduces flexibility as well as programmability in 5G networks by allowing *(i)* to have different features for one or more network functions according to the service to be provided (e.g., different mobility management procedures for different types of traffic) and *(ii)* to re-design network functions without the need of deploying new hardware

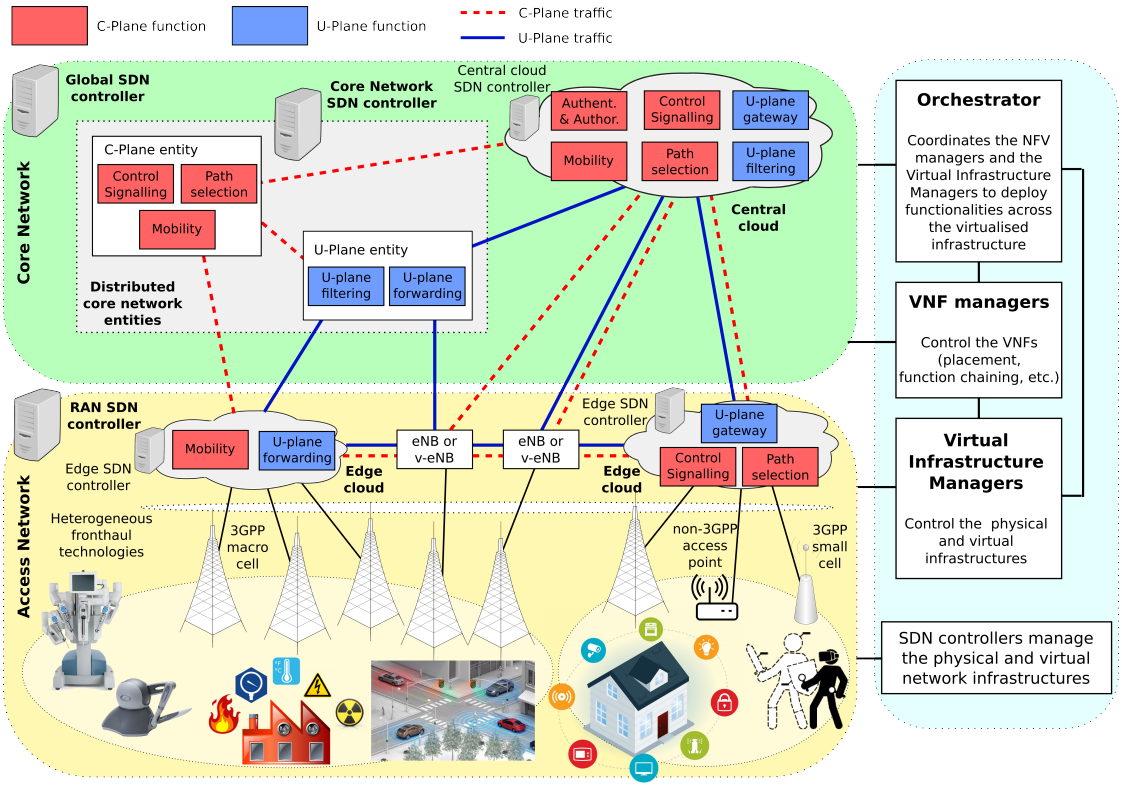

Fig. 2. The architecture of 5G mobile networks enhanced by softwarization and virtualization. The drawing depicts some examples of 5G applications as discussed in Sec. III-A and also shows some examples of network functions and related placements from works [4], [75]

entities to run such functions. Virtualization, on the other hand, increases the level of flexibility in the network by running network functions in a virtual environment and allow for them to run in different locations. Thus, by exploiting NFV, network functions can be placed across the network according to the needs of the service to be deployed or the status (congestion, etc.) of the network. In addition, the exploitation of SDN to manage the network topology (e.g., link/path configuration), increases the flexibility of the network by means of having a network architecture which varies based on the placement of network functions.

These aspects are highlighted in Fig. 2, which shows that 5G architecture will be composed of a set of U/C-plane functions implemented as Virtual Network Function (VNF)s that can be placed in different ways within the network according to the needs of the traffic and the computation capabilities available at the edge.

To summarize, flexibility and programmability can thus satisfy the following requirements to meet different applications needs: *(i)* facilitating network re-configuration to react to new

requests or traffic changes; *(ii)* resiliency; *(iii)* supporting dynamic network topologies; *(iv)* enabling service-aware QoS.

2) Scalability: Scalability is another key feature to be introduced in 5G networks and allow the support of different business applications and huge amount of devices on the same network. To introduce scalability, different aspects need to be considered. First of all, as highlighted in Sec. II-C2, current 4G networks rely on the fact that the information along the network, even in local or short range communications, travels all the way to the PGW to be routed to the receiver with consequent scalability issues at the PGW. An effective way to introduce scalability in 5G networks is thus to bring some of the network functions closer to the RAN, with additional benefits in reducing delay and control overhead. By softwarizing network functions, these can be designed in a modular way, thus splitting the functionalities of the PGW into sub-functions that could be placed through NFV at the edge if the service to be supported requires low-latency QoS. This is highlighted in Fig. 2, where functionalities relevant to U/C-planes can be placed into edge clouds to cut delays and overhead for some specific services.

Another constraint for extra scalability in 4G is how GTP tunnels are handled. As previously discussed, GTP tunnels add complexity, reduce network scalability and represent an obstacle to the infrastructure convergence. In this sense, 5G architecture needs to enable interoperability between different communication systems and co-existence between different mediums, technologies and protocols. As well, 5G networks must be able to easy integrate different wireless technologies into the mobile core, and also provide a reliable backhaul convergence with fixed networks. In terms of integration and interoperability, 5G needs to facilitate seamless plug-and-play of devices using generic and standardized IoT protocols. In this direction, virtualization can introduce scalability as follows. Since VMs are going to handle most of the control signalling independently for each service, the requirements for a central control entity to handle simultaneously a high number of users or devices can be relaxed. As seen, not all the applications involve the same requirements; this feature could be used to the advantage of improving the scalability. For instance, mobility-related functions can be placed in different location of the network according to the features of the devices belonging to different services (e.g., different placement for functions handling mobility in vehicular and static scenarios).

An example is provided in Fig. 2, where *(i)* different edge clouds have different U/C-plane capabilities in order to scale the network according to their needs and *(ii)* core network functionalities can be placed in either distributed entities or central core clouds according to the needs

of the provided services.

To summarize, scalability can handle the following requirements of 5G applications: *(i)* providing low end-to-end latency; *(ii)* network availability; *(iii)*; supporting a massive number of devices; *(iv)* supporting an easy integration and interoperability with other technologies and protocols.

3) Re-designed management of U/C-planes: The applications discussed in Sec. III-A are heterogeneous in terms of requested QoS. As handling QoS involves both C- and U-plane operations, such planes needs to be re-designed in order to allow an easier and more efficient traffic management.

A first aspect to consider is that the majority of the procedures in current 4G networks are triggered by network entities rather than the UE. This aspect is controversial when incorporating low-latency services as it adds a lot of complexity when performing active QoS management, based on instantaneous decision making (i.e., congestion, change of traffic prioritization, etc.). The reduction of signalling is studied in [76], which considers assistance from the UE to reduce signalling required to select specific network functions. As the impossibility to provide instant QoS may seriously impair the ultra-high reliability requirement for some of 5G applications, UE-triggered signalling procedures should be designed in 5G networks. The work in [77] represents a first example towards this direction, where SDN procedures are extended to reach wireless terminals, thus allowing mobile terminals to assist the network decisions and policy enforcement by providing their network connectivity knowledge such as detected cells and related link qualities.

Going one step further, it is necessary to integrate service specific signalling to enable, for instance, enhanced MTC communications asking for very low overhead or very low latency or fine-granular signalling for location-based services. In this case, virtualization allows to have different sequences of functions (e.g., SFCs) designed with different granularity and thus with different overhead and latency features. From Fig. 2, we can highlight that low overhead communications can be enabled for instance by reducing the number of VNFs involved in the management of a given traffic according to the requirements and the features of the target service.

To summarize, a proper re-designed management of U/C-planes in 5G networks can satisfy the following requirements: *(i)* supporting ultra-low latency; *(ii)* providing active QoS management; *(iii)* guaranteeing reliability; *(iv)* supporting UE-triggered signalling.

TABLE II

MAPPING BETWEEN CAPABILITIES, REQUIREMENTS OF USE CASES AND SOLUTIONS OF 5G NETWORKS

5G Capabilities	5G Use Case Requirement	5G Solutions	Role of SDN & NFV
Flexibility & Programmability	• Fast reconfiguration • Resiliency • Dynamic network topology • Service aware QoS	• Network slicing • Service-based network functions • Service-based network topology	• NFV manages the functions and the SFCs of each slice • NFV manages the function placement according to the QoS needs of the slices • SDN manages the network topology (link/path configuration) of each slice • SDN provides traffic isolation among slices
Scalability	• End-to-end low latency • Availability of the Network • Massive number of devices • Easy integration and interoperability with other technologies and protocols	• Cloud-RAN • Mobile edge • Network functions closer to the UE • Services closer to the UE • Dedicated control signal management based on service requirements	• NFV moves network functions closer to the edge • SDN updates the topology to handle function relocation
Management of U/C-planes	• Ultra-low latency • Active QoS management • Reliability • UE triggered signalling	• Slice-based C-plane functions • Slice-based U-plane functions • C/U-plane decoupling • DUDe • Multi-connectivity	• NFV provides ad-hoc C/U-plane functions for each slice • NFV handles function placement according to the DL/UL and C/U-plane configuration • SDN updates the network topology according to the DL/UL and C/U-plane configuration

C. 5G solutions

The introduction of the capabilities discussed in Sec. III-B requires radical changes in the network architecture. The solutions under consideration to implement above discussed capabilities in 5G networks are listed in the following of this Section and summarized in Tab. II.

1) Network slicing: Handling different verticals with different application environments, ranging from broadband services to criticalapplications such as Industrial networks [78], requires solutions that are custom-designed for each scenario/traffic to be delivered; this leads to the concept of *network slicing* [4]. According to the Next Generation Mobile Networks (NGMN) alliance, network slicing primarily targets a partition of the core network, and it is the collection of 5G network functions that are efficiently combined to satisfy one specific business use case, and avoid all unnecessary functionalities [79].

Network slicing helps to introduce flexibility. Instead of having only one core network with one set of functionalities handling all the services provided of network, 5G network will be flexible as each service would have a different dedicated core network slice created that ensures the QoS accomplishment. Slicing also means having ad-hoc U/C-plane functionalities for each business use case. This will further allow to increase the scalability, to reduce the signalling overhead and latency by means of avoiding all unnecessary functionalities.

According to the study done by 3GPP in [80], network slicing is presented as a valid solution to simultaneously handle multiple verticals in 5G networks in a robust way. Dedicated core networks are studied and evaluated in [81], [76]. In particular the 3GPP has introduced the architectural enhancements required to support dedicated core networks, and covers the assignment and maintaining the association in mobility.

Softwarization and virtualization help to bring into reality the idea behind slicing. Indeed, while in 4G networks the creation of dedicated logical networks would require novel ad-hoc network entities and consequently high cost, softwarization and virtualization allow to slice the physical network with considerable reductions from a CAPEX and OPEX point of view: this is due to the fact that the introduction of a new slice involves only the introduction of additional network functions (i.e., software) within the network instead of new infrastructure. In the context of providing dedicated logical networks with specific functionalities, the 3GPP has recognised the use of virtualization as a high level architectural requirement in [74]. Network slicing has numerous business impacts on the operation of the network, and the way resources are shared

between different applications and verticals [82], [83]. It also requires software-based functions that can be chained and moved across the network with a certain degree of freedom; in this sense, NFV can help to build and place VNFs and SFCs a slice is composed of. As a second aspect, SDN is needed to create different traffic paths for the deployed slices as each slice could need a different topology to meet its own QoS target.

2) Towards the edge: Supporting use cases with stringent QoS needs in terms of latency or managing services asking for low overhead require to bring functionalities closer to the edge [84]; this would also increase the scalability of the network in addition to the provisioning of a better QoS.

To this aim, the Cloud/Centralized RAN (C-RAN) [85] paradigm is gaining importance in the 5G architecture. In 5G networks, C-RAN is moving from the concept of having a central manager of digital function unit, a.k.a. baseband processing unit (BBU), to a more general concept of function split [86], [87]. To further increase the scalability, to reduce the complexity and to improve the QoS, 5G networks are moving towards a mobile edge computing approach where services (such as caching) in addition to functions are moved closer to the edge [88].

When considering edge-oriented deployments such as C-RAN or mobile edge, the role of virtualization and softwarization is thus to re-configure the network by moving network functions or services and to properly update the related traffic paths [89]. In case a slice's QoS is deteriorating or the traffic demand from this slice is overloading a given location of the network, NFV can trigger a re-location of network functions closer to the edge and SDN can update the network topology to react to the above mentioned changes. In the field of bringing virtualization closer to the edge, the virtualization of BS's low level functions is addressed in [90]. The main challenge identified in [90] is the virtualization of compute-intensive baseband functions such as the PHY layer, typically implemented on dedicated hardware or on general purpose hardware accelerators. This use case of physical layer virtualization is discussed in terms of acceleration technologies.

3) Decoupling: In the direction of re-designing C/U-planes to make them more efficient and support increased capacity, 5G is introducing radical changes in some of basic features of mobile networks. From this point of view, a disruptive concept is the *decoupling* which can be observed from different point of view.

A first decoupling aspect is related to the cell association process, which involves having a separate cell associations for uplink and downlink, a.k.a. downlink and uplink decoupling

TABLE III

SUMMARY OF STANDARDS CONTRIBUTIONS TO VIRTUALIZATION IN 5G NETWORKS

Standardisation Body	Reference	Scope
ITU-R	• M.2083-0 [91]	• Define the framework and overall objectives of IMT-2020 • Flexible network nodes, highly configurable based on SDN and NFV
3GPP	• TR 22.891 [80] • TR 23.799 [74] • TR 23.707 [81] • TR 23.711 [76]	• Network slicing as key issue to deliver services to industry verticals • High level requirement of virtualization for next generation network architecture • Dedicated core network enhancements for different services • Enhancements on dedicated network selection to reduce signalling, allow for UE assisting information
ETSI NFV	• GS NFV 001 [92] • GS NFV-IFA 001 [90]	• NFV potential use cases • Mobile core virtualization advantages and challenges • Layer 1 virtualization with acceleration technologies

(DUDe) [93], [94]. By allowing users and devices to have different access points for the different traffic directions, the capacity of the network can be maximised (for instance, by having a downlink connection with a macro BS while the uplink one is carried trough a small cell). To enable this solution, a shared Medium Access Control (MAC) layer among involved cells would be needed to handle control procedures such as Hybrid Automatic Repeat Request (HARQ) acknowledgements. In addition, the Radio Resource Control (RRC) layer needs to be centralised and shared among serving cells, since parallel RRC connections would add too much complexity in the UE side [95].

Another decoupling aspect is having a decoupled C/U-plane association to increase for instance the reliability of C- and U-plane traffic or to reduce the amount of control traffic (for instance, in an Intelligent Mobility (IM)/ITS scenario, involving having a C-plane connection to a macro BS to minimize the number of handover while having a U-plane connection with small cells to increase the capacity) [96].

Above considered solutions fall in the multi-connectivity philosophy of 5G networks, where a device shares resources with multiple BSs [97]. Virtualization and softwarization are thus needed

to manage above presented scenarios by migrating network functions (for instance, MAC/RRC entities) to guarantee the availability and the continuity of C/U-plane anchor points when the DL/UL or C/U association changes.

IV. Towards SDN & NFV in 5G: state of the art

In this Section we will provide a detailed overview of the works in literature that focus on virtualization and softwarization for 5G systems. We summarize the lessons learned by this review by highlighting the benefits of the proposals in literature and linking them to the requirements of 5G systems highlighted in Table IV.

A. Design of network architecture

The first aspect to be considered when analysing virtualization and softwarization in 5G systems is related to the drastic changes in the network architecture. The proposals in literature in terms of architecture design can be classified into three different categories, such that each category is characterized by placement of network functionalities.

1) Full-Cloud Migration Architecture: This solution deals with the placement of all network functionalities in an operator-owned cloud, which centrally runs all network functionalities and signalling. The full-cloud migration brings benefits in terms of network availability, easy integration, and interoperability with other technologies and protocols; these can be integrated at the cloud-level.

An example is shown in Fig. 2, where we can note that eNodeBs (or virtual-eNodeBs) can be directly attached to operator-controller central cloud. The full-cloud migration architecture is considered in the following works [75], [49], [98], [99], [17], [100], [101]. The idea at the basis of the full-cloud migration is that all the decisions in terms of U/C-plane management are centrally performed at the cloud-side.

The central solution allows cost savings (as it avoids the use of computation and storage capabilities geographically distributed) and easy maintenance and updates while guaranteeing high computational and storage capabilities [75]. Basta et al. propose the exploitation of SDN to handle incoming/outgoing signalling and data traffic between the BSs and the cloud as well as within the cloud (i.e., intra-cloud communications) [75] . Works in [100], [101] propose additional details on this architecture by defining three level of controllers. The *device controller* is in charge for access network selection, the *edge controller* handles authentication, security, admission control,

routing, handover, while the *orchestration controller* performs network management in terms of C-plane instantiation/management and U-plane balancing.

The disadvantages of the full-cloud migration architecture are in terms of cloud infrastructure performance. Indeed, due to frequent operations in the C-plane [47], [54] as well as the large amount of traffic to be managed by the PGW [49]), the eNodeB-to-cloud connectivity in terms reliability, capacity, latency, etc., could become a bottleneck reducing the advantages of this architecture. Another drawback of this solution is the limited flexibility. Indeed, as all the functions of the mobile core are centralized in the cloud, the only flexibility enabled by this architecture is in terms of different function chaining within the cloud and not in terms of different function placements within the network.

2) C-plane Cloud Migration Architecture: This architecture deals with the functionality split of U/C-planes, as considered in [3], [75] and observing from Fig. 2, the eNodeBs (or virtual-eNodeBs) are managed by a central cloud from a C-plane point of view, while from a U-plane point of view they are attached to U-plane entities distributed across the network.

The C-plane cloud migration architecture allows higher flexibility by migrating the virtualized U-plane components according to the traffic load or service requirements [60]; the bottleneck effect of a full-centralized mobile core in the cloud is thus reduced. The negative aspects are related to the need of moving critical functions (e.g., GTP tunnelling) on U-plane elements to avoid forwarding all data packets to the cloud. This could require the use of customized hardware/middleboxes or supplying switching elements with programmable platform to handle such functions.

3) Scenario-based Migration Architecture: The principal idea of this solution is to deploy network functionalities in both the cloud and mobile infrastructure; the functionalities are migrated according to overload conditions and service requirements [75], [102], [103]. As seen in Fig. 2, some network functionalities can run on edge clouds [104], to achieve a higher flexibility and scalability. This architecture allows functions to be instantiatied at the edge for services dealing with delay-sensitive applications or for scenarios requiring additional functionalities (e.g., the support of non-3GPP access points).

Yousaf et al. [102] analyse the advantages of moving functionalities across the mobile infrastructure, and show that up to 48% of network resources can be saved when moving functionalities close to the radio access. Another example can be found in [75], where Basta et al. propose to move charging functions in the cloud, with statistics exchanged with other entities of the

mobile infrastructure. Haneul et al. [103] propose a model to design the optimal placement of functionalities in order to minimize the packet transmission cost. Such a cost takes into consideration the expected residence time of a function for each candidate location and the cost associated to the selected location.

The scenario based migration architecture has the following benefits: *(i)* fast reconfiguration and improved resilience in case of network changes or failures by means of supporting a dynamic network topology; *(ii)* potentials for low end-to-end delay and ultra-low latency; *(iii)* support of massive number of devices. On the cons side, this solution requires state synchronization and orchestration between the network clouds and the entities distributed in the mobile network: this could introduce additional signalling overhead [75]. In addition, when moving functionalities out from the central cloud, the disadvantages are in terms of a quite minimal exploitation of processing and storing capabilities of a cloud framework as well as the global management of the network.

B. Virtualization of network functions

The architectures discussed in Sec. IV-A and shown in Fig. 2 are based on the idea that network functions can be placed and migrated across the network according to the needs or the network operator and/or the traffic to be managed. As a consequence, this dictates for the need of identifying the network functionalities of relevance for each network function of U/C-planes. In addition, the needs of placing/migrating functions dictates for network functions to be decoupled from the underlying hardware.

Einsiedler et al. [4] define the following set of functionalities to be virtualized: *(i)* authentication and authorization; *(ii)* addressing of devices and entities; *(iii)* forwarding path management; *(iv)* mobility management; *(v)* context aware engine, which collects information from each entity to achieve an overall overview of the network status; *(vi)* optimization function, which evaluates when and how to react to network changes. The work in [4] takes into consideration different use cases (ranging from static sensor networks to mobile broadband services and automotive and transportation services) with the aim to highlight different functionalities needed by different use cases. For example, static environments do not require mobility or optimization functions, these are necessary functionalities when considering transportation systems. This work provides, thus, a first example underling how NFV can enable flexibility in 5G systems to support network

slicing by considering the placement of different VNFs according to the needs of the enabled slices.

More recent studies also focused on defining functionalities of different network entities (i.e., SGW, PGW, MME) and addressing possible solutions for implementing such functions in the virtualized environment of 5G mobile core.

1) SGW's and PGW's functionalities: Basta et al. [75] present a classification of SGW's and PGW's functions by taking into consideration the impact on U/C-planes. The functionalities of SGW and PGW are classified in [75] as follows:

- *control signalling*, which receives and triggers messages and calls the corresponding functions;
- *resource management logic*, which maps the bearer quality attributes according to the QoS requirements expressed in terms of Quality Class Identifier (QCI);
- *U-plane forwarding rules*, defined for each S1-U and S5/S8 bearer established;
- *U-plane forwarding*, i.e., the switching fabric or hardware that carries out the data flows processing;
- *GTP matching*, i.e., GTP headers' matching, encapsulation and decapsulation for both data and signalling traffic;
- *U-plane filtering and classification*, which is handled by the PGW to identify packets based on users' profiles and policies for uplink and downlink data traffic;
- *charging control*, which takes place in the PCRF within the PGW and can be classified into *offline* (based on charging data records) and *online* (triggered by the U-plane traffic) charging.

The work in [75] proposes an SDN realization of the above considered function classification in order to guarantee the support of a dynamic network topology; [75] evaluates the related API's capabilities (i.e., southbound APIs). Concerning the C-plane related functions, authors argue that it is relatively straightforward to integrate and comply with the OpenFlow controller framework. In detail, they propose to map the resource management logic to the centralized switching module, which also needs to be enhanced to include user profiles and policies to be exploited for resource management decisions; this allows to support service-aware and active QoS management. Sama et al. [105] propose to use a module on top of the SDN controller to run control functions of SGW and PGW by using REST APIs as northbound interface.

Focusing on the U-plane forwarding rules, these do not pose any additional effort on OpenFlow

according to [75]. On the contrary, GTP tunneling becomes a challenging aspect as the current production version 1.3.1 of OpenFlow is limited to the matching of layer 2/3 with addition fields such as TCP/UDP ports [106] and thus does not support GTP header matching. To overcome this limitation, [75] proposes different strategies. The first one is to implement a module within the controller to handle GTP functions. Although this solutions matches the current OpenFlow implementation, it requires the transmission of each packet to the controller, with consequent additional overhead and latency. Other strategies could be the deployment of middleboxes, or the enhancement of OpenFlow switches as well as supplying switching elements with programmable platform that should enable the provisioning of enhanced functions such as GTP tunnelling. Concerning GTP tunnelling, [105] proposes to use GTP tunnels only between the eNodeBs and the first entry point of the mobile core (which is supposed to be an OpenFlow-enabled switch) instead of having tunnels up to the PGW.

Basta et al. [75] propose to provide U-plane filtering and classification through the use of matching rules corresponding to the flow filters (source and destination IP addresses, source and destination ports, protocol ID) specified in [107]. Finally, charging functions are proposed to be implemented with the use of counters and statistics exchanged between the switches and the controller. Offline charging can be handled by using an additional module to collect the charging data records based on the OpenFlow counters, while the online charging becomes challenging because OpenFlow switches do not keep state of the forwarded data flow. Again, the use of middleboxes, enhancing OpenFlow switches or using programmable platform can overcome this limitation.

Similarly to [75], Sama et al. [62] propose to replace the control protocols that run on S1-MME (i.e., between the MME and the eNodeB) and S11 (i.e., between the MME and the SGW) interfaces by the OpenFlow protocol. In [62], the OpenFlow controller manages the forwarding plane between the eNodeB and an advanced switch able to encapsulated/decapsulate GTP packets (namely the SGW-D) and between the SGW-D and the PGW. Authors define the SGW-C, i.e., the intelligence of the SGW that is responsible to establish the GTP tunnels; this entity is centralized and runs on top of the OpenFlow controller.

A totally different approach is instead considered by Jin et al. [108], where the PGW is replaced with a simple switch and a middlebox managed through SDN. In this proposal, in order to support low end-to-end delay, ultra-low latency and a massive number of UEs, all packet classification are performed at the access edge, using software switches along with local

software controllers. The proposed SDN controller implements the LTE signalling protocols used between UEs and the MME for the sake of connection establishment and disconnection, tracking area update, paging, etc.

2) Mobility management: Mobility management in current 4G deployment involves high delay up to 0.7s for intra-LTE handover [109], while this delay can also reach 2 minutes when considering handover between 3G and 4G systems[9]. 5G applications push strict requirements in terms of handover delay, and this dictates for a novel design of mobility management strategies.

The research work in [110], [111] analyse the handover mechanism of LTE networks and underlines that the majority of the signalling messages of this procedure are communicated over the backhaul segment of the network, i.e., the links between the eNodeBs and the MME. This is due to the fact that the MME has to request a new connection of the target eNodeB in order to perform the handover; this procedure has higher overhead in case the serving and the target eNodeBs belong to two different SGWs. The bandwidth of the backhaul comprises the most scarce resources in the mobile network and, consequently, novel handover mechanism should be designed to optimize the utilization of this precious resource. The work in [110] exploits an SDN controller to handle mobility management. In this proposal, the controller manages the eNodeBs in a similar way as switches, i.e., by pushing new forwarding rules for the U-plane. The SDN controller selects the target eNodeB based on a set of predefined policies and information collected by the UEs. From a backhaul point of view, this means that various request/confirm messages between eNodeBs can be avoided and this could potentially bring to a reduction down to 50% of the signalling: this allows a support of massive number of UEs. From a delay point of view, authors in [110] conclude that the overall handover delay can be reduced to two to three times the round-trip time (RTT) between the UE and the SDN controller; this could potentially help to support ultra-low latency applications. At the UE, handover involves energy consumption due to the scanning phase when the UE scans and measures the received signal strength from all the neighboring eNodeBs. The proposal in [110] shows also reduction in consumption as the SDN controller can provide the UE with a list of most suitable eNodeBs to scan.

Morales et al. [112] extend the work in [110] by proposing the handover functionality as a service (FaaS) for 5G systems. [112] proposes a module compliant with the SDN controller managing the SGW and the PGW; such a module allows to maintain an overall overview of

[9]http://www.cellular-news.com/story/46917.php

the UEs, the eNodeBs to which they are connected to and the related SGWs. For this aim, the network topology is handled through a graph that evolves with time to keep trace of users' mobility. Each link between a UE and the related serving eNodeB has a weight that can be used to indicate the status of the UE according to pre-defined thresholds: if the weight is below a given threshold, the SDN controller obtains signal quality information from the eNodeB in order to decide if the handover has to be triggered or not. The graph visualization of the network topology can be easily enhanced by considering additional parameters on the weight calculation (e.g., QoS, backhaul characteristics, load balancing) and by exploiting mobility prediction at the controller.

By extending the work in [110], Duan and Wang [113] propose an authentication handover module (AHM) in the SDN controller, in charge of monitoring and predicting the location of users; accordingly, the AHM prepares the relevant cells before the user arrives. The AHM module is also designed to guarantee seamless handover authentication. In detail, authors aim to modify the authentication procedure for handover, composed of multiple independent verifications [114], with the purpose of designing a monitored seamless procedure without increasing the possibility of impersonation and attacks. Authors propose to use attributes such as identity, location, direction, RTT, and physical layer characteristics as reliable secure context information (SCI) to assist secure handover in SDN-enabled 5G networks by providing an unique fingerprint of the specific device without additional hardware and computation cost [115]. The SCI is stored at the SDN controller, which uses an ascending index to indicate the sequential order of next cells in the UE's moving direction; so doings, the cells visited by the UE need only to check if the SCI created with the information provided by the UE is the same as the one provided by the controller. This approach allows flexibility as the number of attributes associated to the SCI can be tuned according on the security level of the information requested (e.g., lower number of attributes for generic web browsing and higher number of attributes for banking or email services).

The concept of mobility management is extended by the work in [77] where SDN procedures extended to reach and control wireless nodes with the aim to support network-controlled offloading mobility in a multi-access scenario. This approach introduces flexibility for multi-access management supporting offloading without requiring additional protocols to support inter-access mobility.

3) Grouping EPC entities: Hawilo et al. [29] take into consideration the protocol stack with all the related functionalities of the EPC. Authors propose to group together different functionalities to allow an easier instantiation in central cloud environments focusing on supporting massive number of devices and low end-to-end latency; authors argue that it would be beneficial to instantiate each group of functionalities in one physical server depending on the workload. To this end, [29] proposes three different groups of functionalities:

- Migration of the MME with an HSS front-end with reduced functions such as authentication and authorization. This cuts the number of transactions between the MME and the HSS from 1,039,430 [55] to 173,239 per second.
- Migration of the PGW withing the SGW to minimize the number of nodes in the U-plane and reduce to zero the number of transactions between the SGW and the PGW (56,559 per second in 4G, on average [55]).

4) Caching-as-a-service: Due to the exploitation of a cloud environment and virtualization, 5G networks open the opportunity to handle services in addition to network functions, within the network nodes. One example of such services is caching.

To reduce traffic within the core network as well as the network entities along the end-to-end path, caching popular contents at mobile network edges attracted the interest of research community [116], [117]. Following this direction, the work in [118] proposes a novel concept, namely caching-as-a-service (CaaS), dealing with the exploitation of mobile core virtualization to enable virtualized caching inside the cloud center owned by the network operator. The idea is to adaptively create, migrate, scale (up or down), share and release CaaS instances in mobile clouds depending on the user demands and requirements; this could allow to support ultra-low latency and massive number of devices. Authors propose the use of caching VMs that co-exist with other VMs for RAN and CN virtualization. So doing, caching VMs can be migrated from any server to any other ones thus allowing global optimal scheduling while taking into consideration the QoS requirements related to the service requesting the content. Furthermore, any content can be divided into chunks and prepared into packets to be already suitable for delivery. Authors suggest to add a centralized caching controller to manage the available contents and the related caching VMs.

C. Virtual Network Function (VNF) Placement

With the introduction of VNFs, there is a new degree of freedom in optimizing networks, by moving the functionalities across different entities, ranging from central cloud, to edge cloud or even distributed across the core network. Works in [75], [119], [120] focused on different placement strategies for the SGW and PGW, with the main target of supporting low end-to-end latency.

Basta et al. [75] proposed to take into account the following parameters when considering the placement of functionalities: costs, data overhead, end-to-end delay. By considering data overhead, this approach could enable a service-aware QoS management. Focusing on the functionalities of SGW and the PGW, [75] argues that the advantages and disadvantages of function placement into the central cloud can be summarized as follows: *(i)* in terms of costs, having all the functions to the cloud would introduce a drastic cost reduction; *(ii)* in terms of data overhead, signalling functions into the cloud would not impact significantly the performance, while placing other functions into the cloud would increase drastically the data overhead; *(iii)* in terms of end-to-end delay, having signalling, resource management logic and filtering into the cloud would not impact significantly the performance, while placing into the cloud charging and GTP tunnelling would involve a drastic delay increase.

In the field of function placement, Taleb et al. [119] have devised a set of solutions to the problem of VNFs placement on federated cloud to create efficient virtual network infrastructure. Authors focused on the placement of functionalities by considering two conflicting aspects: *(i)* insuring QoS to UEs by placing data anchor gateways (i.e., PGWs) closer to edge thus enabling service-aware QoS management and potentially ultra-low latency services; *(ii)* avoiding the relocation of mobility anchor gateways (i.e., SGWs) by placing the relevant functionalities far enough from UEs. For this analysis, [119] has considered performance indicators such as SGW relocation cost (in terms of number of messages needed for the sake of serving area update when UEs change their respective SGW), data delivery delay and number of VMs created to run PGWs and SGWs instances.

With the aim to minimize the network load and the delay, Basta et al. [120] have developed a function placement problem able to decide whether to deploy SGWs/PGWs with fully virtualized or decomposed functions according to the mobile core topology and the number of available data centres across the network. Placement of virtual functions however, would not be possible unless

the appropriate mechanism for resource allocation so that VNFs can be chained together and form SFC. Such resource allocation is studied in [121] where function placement, admission control and embedding in the network infrastructure is modelled as a Mixed Integer Linear Programming (MILP) problem.

D. Mobile virtual networks

White Mobile Virtual Network Operators (MVNOs) have played a role in the arena of mobile networks, the possibility of having different operators which share the same physical infrastructure is becoming more important in the ecosystem. The reason behind network sharing is that a large portion of sites are under-utilized: in [122] it was reported that around 20% of all sites carry about 50% of total traffic. This means that numerous sites carry a negligible level of traffic, although their consumption in terms of energy and computational resources is high. A market survey done in 2010 [6] showed that over 65% of mobile network operator in Europe were deploying different types of network sharing solutions. According to the survey, these solutions save up to 40% in terms of CAPEX and up to 15% in terms of OPEX over a five year period.

1) Network sharing: Network sharing solutions are already available and standardized by 3GPP [123] in order to improve network availability, integration and reliability. As discussed in [124], [125], network sharing can be passive and active, where the former refers to the reuse of physical sites, tower masts, cabling, cabinets, power supply, air-conditioning, etc., while the latter refers to the reuse of backhaul, base stations, and antenna systems. Studies show that operators could save at least 40% of their costs with active network sharing[10].

Above considered aspects motivate the growth of a novel type of operators, referred to as mobile virtual network operator (MVNO), that see in network virtualization a viable solution to implement network sharing. To enable end-to-end cellular network virtualization, the mobile CN and RAN have to be virtualized, where the latter involves specific problems due to the characteristics of wireless access links (e.g., varying channel conditions) [126]. A detailed survey on the issues in implementing MVNO can be found in [125], [124], and references therein. Costa-Perez et al. in [124] consider the work by 3GPP RAN Sharing Enhancements (RSE) of

[10]http://www.cellular-news.com/story/36831.php

the System Architecture Working Group 1 (SA1) [127] and identify the objectives of RAN sharing in terms of:

- *isolation*, i.e., the reserved resource allocations of one entity is not affected by load variations and physical fluctuations of other entities;
- *customization*, i.e., each entity has the flexibility to program its virtual base station to optimize its service delivery;
- *utilization*, the overall utilization of base station's resources is maximized by allowing to use unused resources of one entity across other entities.

Khan et al. in [125] introduce a network configuration platform to decouple the physical infrastructure (advanced mobile access, optical mobile network and service delivery network) from the users of the infrastructure (i.e., the MVNOs). In detail, [125] proposes a network architecture where the physical network infrastructure is shared by MVNOs while management and control of the physical infrastructure, virtual network creation and maintenance, and the usage of the virtual networks are decoupled from each other, to reduce management complexity in both RAN and core networks. The [125] summarizes open issues for MVNOs as follows:

- *Virtualization and its cost.* Virtualization can be applied at different layers of the protocol stack, each one having its own characteristics in terms of performance metrics and cost of implementation and maintenance [128]. If the baseband physical layer functions are softwarized and run as VMs, solutions for independent hardware acceleration are required to guarantee timely completion of CPU jobs. Functionalities that can be more directly virtualised are for example the RRC or Packet Data Convergence Protocol (PDCP) layers, as they do not require strict synchronization. This will influence the functionalities to be transferred from the infrastructure providers to MVNOs.
- *Interfaces.* New interfaces need to be defined and standardized for the purpose of representation, request, and reservation of virtual resources (e.g., through means of a resource description language).
- *Unified C-plane.* A challenging task is to unify the control of the network, due to the fact that each MVNO may operate its own control protocol. As a consequence, the control/management of the network has to interact with all these protocols and to harmonize their functionalities.
- *Inter-MVNO management.* MVNOs are isolated from each other from a both data and control

point of view. This involves challenges in managing inter-MVNO functionalities, such as roaming, which needs to be handled by the central management of the shared network.

2) BS virtualization: An effective RAN virtualization is achieved through the BS virtualization. i.e., different virtual networks (referred to as *slices* in [124], [129]) managing different sets of flows that share a physical BS. This could allow to support a fast re-configuration of BS functionalities, as well as an easy integration and interoperability.

The BS virtualization can be performed at either the hardware or flow level. Solutions to implement the first approach, involving the use of dedicated spectrum, are already available. An example is the virtual base transceiver station (vBTS)[11], able to run multiple base station protocol stacks in software.

The flow level virtualization, which is instead based on the idea of using shared spectrum, represents a more viable solution to save costs and to accelerate network rollouts by achieving effective resource multiplexing and by allowing deployment scenarios where the virtual networks do not own spectrum. To this aim, Zaki et al. [130] propose to implement a *hypervisor* in the eNodeB to manage different entities (each corresponding to a virtual eNodeB managed by a different MVNO) running in different virtual machines within the same eNodeB. The entities provide inputs, which can be either fixed or dynamic, to the hypervisor in terms of user channel conditions, loads, priorities, QoS requirements as well as information related to the contract of each of the virtual operators. Thus, the hypervisor allocates spectrum to the different entities to accommodate their requests while satisfying the specified guarantee. The proposal in [130] only ensures isolation of resources and fails to provide customization of resource allocation across their users.

Tseliou et al. in [131], [132] proposes the resources negotiation for network virtualization (RENEV) scheme that allows slicing and on demand by transferring physical resources among multiple virtual BSs. The aim of RENEV is thus to move resources not utilized by a given virtual BS to another virtual BS in order to improve spectrum utilization and performance of UEs attached to virtual BSs. Costa-Perez et al. [124] discuss a hierarchical architecture, namely the network virtualization substrate (NVS) [133], where a *slice scheduler* ensures isolation across slices while a *flow scheduler* enables flexibility of flow scheduling to the different slices. To integrate NVS into the current BS implementation, [124] proposes to tag each flow with

[11] Vanu Networks, http://www.vanu.com/

an additional slice ID. Therefore, in every scheduling frame, only those flows belong to the slices that are served in that specific frame, will be scheduled. Focusing on the slice scheduler proposed in [124], NVS defines a resource- or a bandwidth-based reservation approach, aiming to perform the resource allocation in terms of a fraction of the total BS's resources or in terms of aggregate bandwidth, respectively. The [124] also defines a tunable utility function, which takes into account the resources allocated to the slices in terms of bandwidth or physical resources. Such a function is maximized in order to maximize the overall utility of the BS, while ensuring slices' requirements according to the service level agreement (SLA) between the slice owner and the physical network owner. The objective function defined by NVS is a weighted log-utility function that provides fairness among slices (similar to proportional fairness after meeting the reserved bandwidth or resources for each slice). It is worth noting that the utility function defined in [124] does not focus only on the data traffic. Indeed, when updating the cumulative resources or bandwidth of a slice, NVS also accounts for the cumulative control signaling generated by the users' within the slice. This represents a key design feature to provide isolation of resources against slices that generate high amount of signaling. Concerning the flow scheduler, NVS defines three different modes: scheduler selection, model specification and virtual-time tagging. In the scheduler selection, a slice can choose from a set of common flow schedulers already implemented within the BS. In the model specification mode, each slice provides a weight distribution function (taking into account different parameters such as average rate or channel conditions) to schedule the flows within its own slice. In this mode, NVS performs the flow scheduling for each slice by selecting the flow with the smallest flow, and thus updating the weights of all flows. In the virtual-time tagging mode, each slice implements its own scheduler and NVS provides real-time, flow-level, feedback to each slice.

3) BS virtualization with multi-RAT capabilities: Extending the work in [124], Rahman et al. [129] propose an architecture for supporting multiple radio access technologies (RATs) in a virtualized BS through the use of a hypervisor composed of: slice manager, resource controller (in charge of physical resources such as computation and storage), spectrum manager (which virtualizes the air interface). The hypervisor interacts with a single radio controller (SRC), which is a module with multi-RAT capabilities. This solution extends the resiliency of BS virtualization.

4) Multi-tenancy: 5G network is expected to handle the co-existence and cooperation of multiple tenants that require the same network functionalities. This represents the so-called multi-tenancy environment, which refers to the possibility that a single instance of a software

application may serve multiple tenants of the network [134].

Multi-tenancy in 5G systems has recently attracted the interest of research community. Condoluci et al. [67] investigate multi-tenancy in smart city environments by focusing on the management of the management of emergency services. In [67], possible services (e.g., transport services, roadside equipments, officers, traffic control) that could be enabled in a 5G smart city scenario are identified and modeled as tenants. The interactions between these tenants are then analyzed in order to define the inter-tenant communications for handling emergency services. Authors propose to exploit the SDN paradigm to manage the QoS of inter-tenant communications. By having a global view of the network, an SDN controller is able to have visibility of the traffic within the network and to reserve the resources for a low-latency management of emergency services.

E. Virtualization of C-RAN

The C-RAN has been widely studied in recent years thanks to its expected benefits in terms of spectrum and energy efficiency as well as CAPEX and OPEX reduction [135]. Softwarization and virtualization are considered as candidate technologies to promote the realization of C-RAN [136], [137], [99], with focus on the design characteristics to virtualize a C-RAN and to enhance its architecture with additional functionalities.

Nikaein et al. [138] present a proof-of-concept prototype for a virtualized C-RAN built upon the OpenAirInterface LTE software implementation [139], Ubuntu 14.04 with low latency kernel (3.17), Linux containers, OpenStack, Heat orchestrator, and Open vSwitch and National Instrument/Ettus USRP B210 RF front-end[12]. As a reference scenario, authors considered an 20 MHz FDD system with single-input single-output and AWGN channel. For this scenario, by analyzing the processing time requirement of 1ms for the LTE sub-frame, the main conclusion is that 2 cores at 3 GHz are needed to handle the total processing of an eNodeB. By considering a fully loaded system, the exploitation of one processor core for the receiver processing assuming 16-QAM on the uplink and 1 core for the transmitter processing assuming 64-QAM on the downlink is able to meet the HARQ deadlines. By comparing different virtualization environments, authors conclude that containers (LXC and Docker) offer near bare metal runtime performance, while preserving the benefits of virtual machines in terms of flexibility, fast deployment, and migration.

[12]https://www.ettus.com/product/details/UB210-KIT

Works in [140], [141], [142] introduce an SDN controller at the network edge to allow traditional BSs to be grouped together to realize a C-RAN in order to support a fast network re-configuration and a dynamic network topology. Such works have the common idea of centralizing some functionalities of the C-plane while allowing some other functions to be implemented at the BS-side. So doing, control decisions that influence several BSs (such as handover, load balancing, transmission power etc.) are made at the central controller, while decisions such as resource allocation are made locally at the BS as they impact less neighbouring BSs; this would allow to better handle QoS. Jin et al. [108] introduce an SDN controller to manage different BSs, while a controller is used for both RAN and CN management.

Lu et al. [59] propose a dynamic allocation scheme of radio resources to deal with the dynamic requests from the virtual networks in order to improve the utilization rate of spectrum managed in the BBU pool. Authors consider the underlying physical resource state information (RSI) and queue status information (QSI) for computing the available resources and the change of business of virtual networks, respectively. Virtual networks are divided into three categories according to their priority: real-time voice business, high-speed multimedia services and the traditional best effort service. Authors propose an enhanced genetic algorithm that completes the resource allocation of the requests from the virtual networks and focus on three different performance rates: resource revenue, request rejection rate and resource utilization. Simulation results show that the proposed algorithm can improve the request rate from the virtual networks significantly, due to the full use of idle and fragment carrier resources; this also allows to improve the spectrum utilization and consequently the infrastructure revenue.

Dawson et al. [143] propose to enhance the traditional C-RAN features by moving a set of CN functionalities within the C-RAN. In detail, authors propose an architecture for a virtual C-RAN composed of three main functions:

- *inter-cell coordination and BBU pool* to manage spectrum resources (examples of a more advanced BBU pool management can be found in [144], [145], [59]);
- *mobility anchoring*, to provide a static endpoint on the U-plane for connections betweens cells belonging to the same C-RAN and this could potentially enable the support of ultra-low latency services;
- *mobility management*, to allow handover between cells belonging to the same C-RAN.

In this architecture, the C-RAN is thus considered to provide a subset of functionalities of the MME and the SGW, thus enhancing the performance in terms of supporting low end-to-end

latency and core network overload.

F. Creation and management of mobile clouds

Mobile cloud in 5G is a promising paradigm to provide computational resources on demand that enables resource-constrained mobile devices to offload their processing and storage requirements to the cloud infrastructure [146]. Nevertheless, the realization of mobile cloud is still challenging, as analysed in [147]. In this field, the enhancements introduced by softwarization and virtualization allow to build mobile clouds in a more efficient way as well as to enable global connectivity across different mobile clouds. The benefits are in terms of supporting dynamic network topologies, enabling low end-to-end latency and ultra-low latency, improving reliability and network availability.

Aissioui et al. [147] consider the idea of exploiting SDN for mobile cloud management. To allow flexibility, this proposal permits to distribute the SDN/OpenFlow control plane on a two-level hierarchical architecture composed of a global controller and several local controllers deployed on-demand, where and when needed, depending on the network dynamics and traffic patterns. While the global controller has an overall overview of the network and performs inter-cloud management, local controllers manage their own clouds. Authors show the benefits of this hierarchical architecture in terms of a reduced number of control messages sent from/to the global controller, which can be easily translated in a higher scalability for the global controller as well as in a reduced delay for mobile cloud management as it is performed closer to the edge. As a consequence, although the proposal in [147] looks close to those considered in [100], [101] for full cloud migrated architectures, this work introduces a novel idea of bringing SDN as close as possible to the edge by enabling UEs to become local controllers which interacts with a central SDN controller.

A similar approach is proposed by Usman et al. [148]. They propose an SDN architecture for supporting device-to-device (D2D) communications to build mobile clouds in scenarios with strict requirements in terms of reliability and ultra-low delay (e.g., public safety). In detail, authors propose an architecture on top of the public safety enhancements for LTE, i.e., proximity services (ProSe) and group call system, which are designed to guarantee inter-operability across different public safety applications. The architecture in [148] associates a D2D controller application to a hierarchy of SDN controllers in the network in order to couple the formation and management of the mobile clouds. The central SDN controller has a global view of all mobile clouds in its

range, while the local controllers (cloud heads) are only aware of UEs in their neighborhood. Compared to [147], [148] brings SDN at the end-device level.

Focusing in detail on the creation process of mobile clouds in [148], this is initiated by a UE that broadcasts a request to nearby devices by using an out-of-band technology (e.g., Wi-Fi Direct, Bluetooth). Once the mobile cloud is formed, it is registered at a central SDN controller; this allows such a controller to have a global view of all the active clouds and the related offered services. The features of the architecture proposed in [148] are:

- scalability, as the central controller receives aggregated information by the cloud head [147];
- spectral efficiency, as LTE links are used to send aggregate data instead of sending small packets from cloud members to the cloud head;
- robustness, as the UEs are still able to communicate with partial support from cellular infrastructure in case of disaster;
- interference reduction, as the central SDN controller has the global overview of the network and it is thus able to exploit solution for global interference reduction (e.g., [149]).

In [148], each UE runs an SDN application that maintains a database of all services and resources that a mobile user is willing to share. This database is shared with the other UEs at the reception of a cloud initialization request, and the aggregated database is thus shared by the cloud head with the central SDN controller. This cloud head is in charge of generating the cloud's authentication key, which is thus transmitted with unicast links by the cloud head to the members of its cloud. The use of a central SDN controller with a global view of the network, also allows to form clouds without involving local controllers with the sake of saving UEs' energy. In addition, the central controller has a global information of all the devices and can thus define a route across the UEs for instance by exploiting the Dijkstra algorithm.

V. Future research

In the previous Section, we reviewed the role of softwarization and virtualization in 5G systems from various angles. In this Section, we derive the research directions that still needs to be investigated to achieve an effective exploitation of softwarization and virtualization in 5G systems.

A first topic to be further investigated deals with the placement of network functionalities. By extending the works in [75], [4], [119], [120], a well-designed cost model should consider as input the target QoS requirements (in terms of data rate, latency, jitter, mobility, priority, etc.)

TABLE IV

SUMMARY OF THE STATE OF THE ART AND MAPPING WITH 5G USE CASE REQUIREMENTS

	Fast reconfiguration	Resiliency	Dynamic network topology	Service-aware QoS	End-to-end low latency	Network availability	Massive number of devices	Integration and interoperability	Ultra-low latency	Active QoS management	Reliability
Network Architecture											
• Full Cloud Migration [75], [49], [98], [99], [17], [100], [101]						✓		✓			
• C-plane Cloud Migration [75], [3]			✓			✓		✓			
• Scenario-based Migration [75], [102], [103], [104]	✓	✓	✓		✓	✓	✓	✓	✓		
VNF											
• SDN realization of SGW and PGW [75], [105]			✓	✓	✓		✓		✓	✓	
• Mobility management [110], [112], [113], [77]	✓	✓	✓				✓	✓	✓		
• Grouping EPC entities [29]					✓				✓		
• Caching as a Service [118]							✓		✓		
SGW/PGW placement											
• Cost, overhead, end-to-end delay [75]				✓	✓						
• QoS and mobility anchor gateway [119]				✓	✓				✓		
• Load and end-to-end delay [120]					✓						
Mobile virtual networks											
• Network sharing [125]						✓		✓			✓
• BS virtualization [130], [124], [133], [131], [132]	✓							✓			
• Multi-RAT BS virtualization [129]	✓	✓						✓			✓
• Service Chaining [121]	✓					✓				✓	
• Multi-tenancy [67]					✓			✓			
C-RAN virtualization											
• SDN-based and enhanced C-RAN [140], [141], [142], [108], [143]	✓		✓	✓	✓				✓		
Mobile clouds											
• SDN for mobile cloud [147], [148]	✓		✓		✓	✓			✓		✓

of the service to be enabled. In addition to the parameters considered in [75], [4], [119], this cost model should take into account the features (such as capacity, latency) of the fronthaul and backhaul links [150] of the network that could drastically influence having functionality in the cloud or in the mobile core. A further aspect to be considered, in the cost model definition, is the differentiation in the available resources in the cloud and in the distributed mobile core. Indeed, the cloud infrastructure is usually characterized by a very large availability in terms of computational and storage resources, while such resources from a mobile core point of view are scarcer and thus more affected by congestion and overload. As a consequence, placing one functionality into the a central cloud does not involve a meaningful reduction in its overall computational capability, while this reduction could not be negligible when considering the allocation of resources across the mobile core architecture. All aspects listed above should be considered in the design of a well defined cost function to model network function placement in 5G network architecture.

The challenges in terms of VNF placement are exacerbated when considering the migration of network functionalities. In addition to the time needed to install one or more VMs needed to migrate one or more VNFs from one location to another, aspects such as security and integrity issues of SDN/NFV environments [151], [152], [153] should be considered to define the overall cost related to function migration. Furthermore, it should be considered that some UEs could exploit U/C-plane splitting or DUDe mechanisms [93], [94] to improve their reliability and/or QoS. This aspect makes more challenging the migration of VNFs due to the latency requirements of some network functionalities (such as HARQ).

Another aspect to be considered deals with the definition of network functionalities. According to the current literature [4], [75], [105], [75], [110], [112], [113], [29] and the recent updates from standardization bodies [91], [80], [74], [81], [76], [90], [92], we are observing a novel trend where network functionalities are no longer considered as a monolithic block implemented in a specific layer of the protocol stack. Recent works [4], [75], [105], [75], [110], [112], [113], [29] highlight different network functionalities of U/C-planes, such as charging, tunnelling, mobility management, handover authentication, etc. Nevertheless, a comprehensive study is still needed to define a full set of functionalities as well as the related *function chaining* [154]. Such a study should be driven by considering aspects such as latency, jitter and data overhead for the communication between network functionalities, to avoid that a fine-granular definition of network functions will involve higher delays and overheads. In addition, APIs for the

communication between network functionalities should be defined to guarantee interoperability.

Another important aspect that is currently not sufficiently investigated in the literature is related to the performance of softwarization and virtualization in 5G environments. Indeed, the importance that such paradigms will have to guarantee flexibility, customization and reconfigurability is well consolidated, but there is a lack in understanding how software-based solutions will be able to guarantee the expected performance in terms of latency and reliability for 5G. Performance degradation when functionalities run in software instead of dedicated hardware is a well-known issue, and this aspect should be adequately considered in the design of 5G network architecture. Works in [138], [139], [90], [92] offer first steps to this direction. Nevertheless, further studies are needed by considering realistic network deployments as well as traffic loads in order to define the virtualization features (orchestrator, real-time hypervisor and kernel, container memory and CPU resources) of the cloud platforms of 5G systems.

A further research topic is related to the effective use of SDN within the network architecture of mobile systems and terminals. Works in [140], [141], [142], [108], [150], [147], [148], [67], [112] consider SDN as a key aspect of 5G architectures and they present the idea of extending the functionalities of SDN controllers by adding capabilities in terms of mobility management and so on. This requires additional studies in order to define how such additional functionalities need to be implemented. A common view is to consider such features as modules on top of the SDN controller. This solution needs to be further investigated by considering different solutions for the northbound interface allowing the communication between such modules and the controller. Sama et al. [105] propose to use the REST API interface for this purpose. Nevertheless, the REST API is a HTTP-based interface, whose applicability in the context of handling huge amount of requests with strict inter-arrival time (as expected for control and signalling functions) still needs to be properly evaluated. In addition, the performance of SDN in large networks still needs to be adequately investigated by taking into account features such as scalability, response time to network changes, overhead, and so on. Finally, the integration of wireless terminals within the SDN/NFV architecture should be further studied. Meneses et al. [77] evaluated the benefits of extending SDN procedures to mobile terminals to support multi-access offloading. Further studies of UE integration within SDN architecture are needed also to better exploit the UE role to optimize network procedures.

VI. Conclusion

The need for flexibility, reconfigurability and customization in 5G mobile networks is pushing vendors and operators to exploit novel paradigms for network deployment. In this direction, softwarization and virtualization are two candidates to play a key role in the design of 5G systems, as they allow to decouple network functionalities from the underlying hardware with consequent benefits in terms of easier network management and upgrade.

This paper aimed to discuss softwarization and virtualization as enablers of 5G architecture deployment. We presented an overview of the evolution of mobile networks from 2G to 4G, highlighting the design drivers behind each generation of the architecture and the novelties that each generation introduced compared to the previous ones. This allowed to obtain a summary of functionalities and entities currently implemented and deployed in mobile systems. We thus illustrated the 5G ecosystem by providing a detailed discussion of its application scenarios. We highlighted the requirements that 5G use cases push on the network architecture and, accordingly, we analysed the limitations of 4G systems in fulfilling such requirements. This allowed to define a set of features to be considered in the design of 5G system architecture, and to define the role of softwarization and virtualization in implementing such features. We considered the recent advances by 3GPP looking at introducing softwarization and virtualization in the standardization of 5G network architecture. We provide a comprehensive overview of the state of the art by surveying the recent proposals in literature focusing on the exploitation of softwarization and virtualization for the implementation of network entities and functionalities of 5G systems. Moreover, we provided a discussion on the lessons learned according to the surveyed literature. Finally, we highlighted the future steps to be considered by the research community to bring into reality a softwarized and virtualized 5G network architecture.

References

[1] J. D. Vriendt, P. Laine, C. Lerouge, and X. Xu, "Mobile network evolution: a revolution on the move," *IEEE Communications Magazine*, vol. 40, pp. 104–111, Apr 2002.

[2] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, pp. 84–91, May 2016.

[3] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *Communications Magazine, IEEE*, vol. 51, pp. 44–53, July 2013.

[4] H.-J. Einsiedler, A. Gavras, P. Sellstedt, R. Aguiar, R. Trivisonno, and D. Lavaux, "System design for 5G converged networks," in *Networks and Communications (EuCNC), 2015 European Conference on*, pp. 391–396, June 2015.

[5] K. Mahmood, T. Mahmoodi, R. Trivisonno, A. Gavras, D. Trossen, and M. Liebsch, "On the integration of verticals through 5g control plane," in *Networks and Communications (EuCNC), 2017 European Conference on*, June 2017.

[6] B. Naudts, M. Kind, F. Westphal, S. Verbrugge, D. Colle, and M. Pickavet, "Techno-economic Analysis of Software Defined Networking as Architecture for the Virtualization of a Mobile Network," in *Software Defined Networking (EWSDN), 2012 European Workshop on*, pp. 67–72, Oct 2012.

[7] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 2347–2376, Fourthquarter 2015.

[8] P. Jain, P. Hedman, and H. Zisimopoulos, "Machine type communications in 3GPP systems," *IEEE Communications Magazine*, vol. 50, pp. 28–35, November 2012.

[9] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT Machine Age with 5G: Machine-Type Multicast Services for Innovative Real-Time Applications," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2016.

[10] M. Dohler and et. al., "Internet of skills, where robotics meets AI, 5G and the Tactile Internet," in *European Conference on Networks and Communications (EuCNC)*, June 2017.

[11] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5g roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17 – 48, 2016.

[12] P. Sidhu, I. Woungang, G. H. S. Carvalho, A. Anpalagan, and S. K. Dhurandher, "An Analysis of Machine-Type-Communication on Human-Type-Communication over Wireless Communication Networks," in *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*, pp. 332–337, March 2015.

[13] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 4–16, First 2014.

[14] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: a survey," *IEEE Communications Magazine*, vol. 51, pp. 148–157, May 2013.

[15] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-Based 5G Networks," *IEEE Wireless Communications*, vol. 23, pp. 56–63, February 2016.

[16] D. Kreutz, F. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, pp. 14–76, Jan 2015.

[17] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *Communications Magazine, IEEE*, vol. 53, pp. 90–97, Feb 2015.

[18] A. Lara, A. Kolasani, and B. Ramamurthy, "Network Innovation using OpenFlow: A Survey," *Communications Surveys Tutorials, IEEE*, vol. 16, pp. 493–512, First 2014.

[19] C. Trois, M. D. D. D. Fabro, L. C. E. de Bona, and M. Martinello, "A Survey on SDN Programming Languages: Towards a Taxonomy," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[20] A. Lara, A. Kolasani, and B. Ramamurthy, "Network Innovation using OpenFlow: A Survey," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 493–512, First 2014.

[21] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A Survey on Software-Defined Networking," *Communications Surveys Tutorials, IEEE*, vol. 17, pp. 27–51, Firstquarter 2015.

[22] B. A. A. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 1617–1634, Third 2014.

[23] F. Hu, Q. Hao, and K. Bao, "A Survey on Software-Defined Network and OpenFlow: From Concept to Implementation," *IEEE Communications Surveys Tutorials*, vol. 16, pp. 2181–2206, Fourthquarter 2014.

[24] ETSI, "Network Function Virtualization: An Introduction, Benefits, Enablers, Challenges, and Call for Action." White Paper, 2012.

[25] ETSI, "Network Functions Virtualisation (NFV); Management and Orchestration." White Paper, 2014.

[26] Q. Duan, Y. Yan, and A. V. Vasilakos, "A Survey on Service-Oriented Network Virtualization Toward Convergence of Networking and Cloud Computing," *IEEE Transactions on Network and Service Management*, vol. 9, pp. 373–392, December 2012.

[27] J. d. J. Gil Herrera and J. F. B. Vega, "Network Functions Virtualization: A Survey," *IEEE Latin America Transactions*, vol. 14, pp. 983–997, Feb 2016.

[28] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 236–262, Firstquarter 2016.

[29] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *Network, IEEE*, vol. 28, pp. 18–26, Nov 2014.

[30] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*. Telecom Publishing, 1992.

[31] G. Gu and G. Peng, "The survey of GSM wireless communication system," in *Computer and Information Application (ICCIA), 2010 International Conference on*, pp. 121–124, Dec 2010.

[32] M. Austin, A. Buckley, C. Coursey, P. Hartman, R. Kobylinski, M. Majmundar, K. Raith, and J. Seymour, "Service and system enhancements for TDMA digital cellular systems," *Personal Communications, IEEE*, vol. 6, pp. 20–33, Jun 1999.

[33] J. Scourias, "Overview of the global system for mobile communications," *University of Waterloo*, vol. 4, 1995.

[34] A. Salkintzis, "A survey of mobile data networks," *Communications Surveys, IEEE*, vol. 2, pp. 2–18, Third 1999.

[35] J. Cai and D. Goodman, "General packet radio service in GSM," *Communications Magazine, IEEE*, vol. 35, pp. 122–131, Oct 1997.

[36] C. Bettstetter, H.-J. Vogel, and J. Eberspacher, "GSM phase 2+ general packet radio service GPRS: Architecture, protocols, and air interface," *Communications Surveys, IEEE*, vol. 2, pp. 2–14, Third 1999.

[37] G. Brasche and B. Walke, "Concepts, services, and protocols of the new GSM phase 2+ general packet radio service," *Communications Magazine, IEEE*, vol. 35, pp. 94–104, Aug 1997.

[38] A. Molisch, *WCDMA/UMTS*, pp. 635–663. Wiley-IEEE Press, 2011.

[39] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, "WCDMA-the radio interface for future mobile multimedia communications," *Vehicular Technology, IEEE Transactions on*, vol. 47, pp. 1105–1118, Nov 1998.

[40] F. Muratore, *UMTS: Mobile Communications for the Future*. New York, NY, USA: John Wiley & Sons, Inc., 2000.

[41] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS: high speed radio access for mobile communications*. John Wiley & Sons, 2007.

[42] A. R. Mishra, *Fundamentals of cellular network planning and optimisation: 2G/2.5 G/3G... evolution to 4G*. John Wiley & Sons, 2004.

[43] Y. Yuan, Z. Zuo, Y. Guan, X. Chen, W. Luo, Q. Bi, P. Chen, and X. She, "LTE-Advanced coverage enhancements," *Communications Magazine, IEEE*, vol. 52, pp. 153–159, October 2014.

[44] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes, "The Evolution to 4G Cellular Systems: LTE-Advanced," *Phys. Commun.*, vol. 3, pp. 217–244, Dec. 2010.

[45] T. Mahmoodi, V. Friderikos, O. Holland, and H. Aghvami, "Balancing sum rate and tcp throughput in ofdma based wireless networks," in *2010 IEEE International Conference on Communications*, pp. 1–6, May 2010.

[46] T. Mahmoodi, V. Friderikos, O. Holland, and H. Aghvami, "Tcp-aware resource allocation in ofdma based wireless networks," in *International Workshop on Cross Layer Design*, pp. 1–5, June 2009.

[47] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.

[48] M. Amani, T. Mahmoodi, M. Tatipamula, and H. Aghvami, "Programmable policies for data offloading in LTE network," in *Communications (ICC), 2014 IEEE International Conference on*, pp. 3154–3159, June 2014.

[49] A. Bradai, K. Singh, T. Ahmed, and T. Rasheed, "Cellular software defined networking: a framework," *Communications Magazine, IEEE*, vol. 53, pp. 36–43, June 2015.

[50] J. Mwangama, N. Ventura, A. Willner, Y. Al-Hazmi, G. Carella, and T. Magedanz, "Towards Mobile Federated Network Operators," in *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pp. 1–6, April 2015.

[51] W. Dong, Z. Ge, and S. Lee, "3G Meets the Internet: Understanding the performance of hierarchical routing in 3G networks," in *Teletraffic Congress (ITC), 2011 23rd International*, pp. 15–22, Sept 2011.

[52] J.-H. Lee, J.-M. Bonnin, P. Seite, and H. Chan, "Distributed IP mobility management from the perspective of the IETF: motivations, requirements, approaches, comparison, and challenges," *Wireless Communications, IEEE*, vol. 20, pp. 159–168, October 2013.

[53] J. Shin, K. Jung, and A. Park, "Design of Session and Bearer Control Signaling in 3GPP LTE System," in *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pp. 1–5, Sept 2008.

[54] B.-J. Kim and P. Henry, "Directions for future cellular mobile network architecture," *First Monday*, vol. 17, no. 12, 2012.

[55] Nokia Siemens Networks, "Signalling is growing 50% faster than data traffic." White Paper, 2011.

[56] W. Zhi and L. Siqi, "Improved scheme of RRC message transmission in LTE system," in *Educational and Information Technology (ICEIT), 2010 International Conference on*, vol. 3, pp. V3–345–V3–348, Sept 2010.

[57] K. Nagaraj and S. Katti, "ProCel: Smart Traffic Handling for a Scalable Software EPC," in *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking*, HotSDN '14, (New York, NY, USA), pp. 43–48, ACM, 2014.

[58] M. Balakrishnan, I. Mohomed, and V. Ramasubramanian, "Where's That Phone?: Geolocating IP Addresses on 3G Networks," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, (New York, NY, USA), pp. 294–300, ACM, 2009.

[59] G. Lu, C. Liu, L. Li, and Q. Yuan, "A Dynamic Allocation Algorithm for Physical Carrier Resource in BBU Pool of Virtualized Wireless Network," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on*, pp. 434–441, Sept 2015.

[60] M. Skulysh and O. Klimovych, "Approach to virtualization of Evolved Packet Core Network Functions," in *Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 2015 13th International Conference The*, pp. 193–195, Feb 2015.

[61] A. Medhat, G. Carella, J. Mwangama, and N. Ventura, "Multi-tenancy for Virtualized Network Functions," in *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pp. 1–6, April 2015.

[62] M. Sama, S. Ben Hadj Said, K. Guillouard, and L. Suciu, "Enabling network programmability in LTE/EPC architecture using OpenFlow," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium on*, pp. 389–396, May 2014.

[63] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selen, and J. Skold, "5G wireless access: requirements and realization," *Communications Magazine, IEEE*, vol. 52, pp. 42–47, December 2014.

[64] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A Survey of Traffic Issues in Machine-to-Machine Communications over LTE," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2016.

[65] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1065–1082, June 2014.

[66] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *Internet of Things Journal, IEEE*, vol. 1, pp. 22–32, Feb 2014.

[67] M. Condoluci, F. Sardis, and T. Mahmoodi, "Softwarization and Virtualization in 5G Networks for Smart Cities," in *EAI International Conference on Cyber Physical Systems, IoT and Sensors Networks*, October 2015.

[68] F. Chiariotti, M. Condoluci, T. Mahmoodi, and A. Zanella, "SymbioCity: Smart cities for smarter networks," *Transaction on Emerging Telecommunication Technologies*, vol. 29, January 2018.

[69] M. D. Cia, F. Mason, D. Peron, F. Chiariotti, M. Polese, T. Mahmoodi, M. Zorzi, and A. Zanella, "Using Smart City Data in 5G Self-Organizing Networks," *IEEE Internet of Things Journal*, vol. 5, pp. 645–654, April 2018.

[70] "5G and e-Health." White Paper, October 2015.

[71] "5G and the Factories of the Future." White Paper, October 2015.

[72] R. S. H. Istepanaian and Y.-T. Zhang, "Guest Editorial Introduction to the Special Section: 4G Health - The Long-Term Evolution of m-Health," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, pp. 1–5, Jan 2012.

[73] M. A. Lema, K. Antonakoglou, F. Sardis, N. Sornkarn, M. Condoluci, T. Mahmoodi, and M. Dohler, "5g case study of internet of skills: Slicing the human senses," in *2017 European Conference on Networks and Communications (EuCNC)*, pp. 1–6, June 2017.

[74] 3GPP, "Study on Architecture for Next Generation System," TR 23.799, 3rd Generation Partnership Project (3GPP), Feb. 2016.

[75] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," in *Future Networks and Services (SDN4FNS), 2013 IEEE SDN for*, pp. 1–7, Nov 2013.

[76] 3GPP, "Study on Dedicated Core Network Enhancements," TR 23.711, 3rd Generation Partnership Project (3GPP), Feb. 2016.

[77] F. Meneses, D. Corujo, C. Guimaraes, and R. L. Aguiar, "Extending sdn to end nodes towards heterogeneous wireless mobility," in *2015 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Dec 2015.

[78] T. Mahmoodi, V. Kulkarni, W. Kellerer, P. Mangan, F. Zeiger, S. Spirou, I. Askoxylakis, X. Vilajosana, H. J. Einsiedler, and J. Quittek, "VirtuWind: virtual and programmable industrial network prototype deployed in operational wind park," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 9, pp. 1281–1288.

[79] "5G White Paper." White Paper, February 2015.

[80] 3GPP, "Study on New Services and Markets Technology Enablers," TR 22.891, 3rd Generation Partnership Project (3GPP), Feb. 2016.

[81] 3GPP, "Architecture enhancements for dedicated core networks; Stage 2," TR 23.707, 3rd Generation Partnership Project (3GPP), Feb. 2016.

[82] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management amp; prioritization in 5g mobile systems," in *22th European Wireless Conference*, pp. 1–6, May 2016.

[83] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing in 5g: An auction-based model," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2017.

[84] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[85] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks. A Technology Overview," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 405–426, Firstquarter 2015.

[86] G. Mountaser, M. L. Rosas, T. Mahmoodi, and M. Dohler, "On the feasibility of mac and phy split in cloud ran," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, March 2017.

[87] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings, "Cloud-ran in support of urllc," in *2017 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Dec 2017.

[88] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, pp. 82–89, Aug 2014.

[89] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud ran," *IEEE Communications Magazine*, vol. 54, pp. 26–32, January 2016.

[90] ETSI, "Network Functions Virtualisation (NFV); Acceleration Technologies; Report on Acceleration Technologies & Use Cases." ETSI GS NFV-IFA 001, Dec 2015.

[91] ITU-R, "IMT Vision Framework and overall objectives of the future development of IMT for 2020 and beyond," Tech. Rep. ITU-R M.2083-0, ITU-R, September 2015.

[92] ETSI, "Network Functions Virtualisation (NFV); Use Cases." ETSI GS NFV 001, Dec 2015.

[93] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Communications Magazine*, vol. 54, pp. 110–117, March 2016.

[94] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Control-Data Separation Architecture for Cellular Radio Access Networks: A Survey and Outlook," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 446–465, Firstquarter 2016.

[95] Z. Corporation, "Comparison between CP solution C1 and C2," Tech. Rep. R2-132383, 3GPP TSG-RAN2 , August 2013.

[96] L. Yan and X. Fang, "Reliability evaluation of 5g c/u-plane decoupled architecture for high-speed railway," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, no. 1, pp. 1–11, 2014.

[97] 3GPP, "Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects." TR 36.842.

[98] P. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *Communications Magazine, IEEE*, vol. 52, pp. 65–75, Nov 2014.

[99] P. Demestichas, A. Georgakopoulos, K. Tsagkaris, and S. Kotrotsos, "Intelligent 5G Networks: Managing 5G Wireless/Mobile Broadband," *Vehicular Technology Magazine, IEEE*, vol. 10, pp. 41–50, Sept 2015.

[100] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and D. Soldani, "Towards zero latency Software Defined 5G Networks," in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pp. 2566–2571, June 2015.

[101] R. Guerzoni, R. Trivisonno, and D. Soldani, "SDN-based architecture and procedures for 5G networks," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, pp. 209–214, Nov 2014.

[102] F. Yousaf, J. Lessmann, P. Loureiro, and S. Schmid, "SoftEPC - Dynamic instantiation of mobile core network entities for efficient resource utilization," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 3602–3606, June 2013.

[103] H. Ko, G. Lee, I. Jang, and S. Pack, "Optimal middlebox function placement in virtualized evolved packet core systems," in *Network Operations and Management Symposium (APNOMS), 2015 17th Asia-Pacific*, pp. 511–514, Aug 2015.

[104] I. Giannoulakis, E. Kafetzakis, G. Xylouris, G. Gardikis, and A. Kourtis, "On the applications of efficient NFV management towards 5G networking," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, pp. 1–5, Nov 2014.

[105] M. Sama, L. Contreras, J. Kaippallimalil, I. Akiyoshi, H. Qian, and H. Ni, "Software-defined control of the virtualized mobile packet core," *Communications Magazine, IEEE*, vol. 53, pp. 107–115, Feb 2015.

[106] ONF, "OpenFlow Switch Specification Version 1.3.1." Technical Specification.

[107] 3GPP, "Policy and charging control architecture." TS 23.203.

[108] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," in

Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT '13, (New York, NY, USA), pp. 163–174, ACM, 2013.

[109] Ericsson, "Voice Handover in LTE Network." White Paper.

[110] T. Mahmoodi and S. Seetharaman, "Traffic Jam: Handling the Increasing Volume of Mobile Data Traffic," *Vehicular Technology Magazine, IEEE*, vol. 9, pp. 56–62, Sept 2014.

[111] T. Mahmoodi and S. Seetharaman, "On Using a SDN-based Control Plane in 5G Mobile Networks," in *Wireless World Research Forum, 32nd Meeting*, May 2014.

[112] A. C. Morales, A. Aijaz, and T. Mahmoodi, "Taming Mobility Management Functions in 5G: Handover Functionality as a Service (FaaS)," in *Global Communications Conference (GLOBECOM), 2015 IEEE*, Dec 2015.

[113] X. Duan and X. Wang, "Authentication handover and privacy protection in 5G hetnets using software-defined networking," *Communications Magazine, IEEE*, vol. 53, pp. 28–35, April 2015.

[114] C.-K. Han and H.-K. Choi, "Security Analysis of Handover Key Management in 4G LTE/SAE Networks," *Mobile Computing, IEEE Transactions on*, vol. 13, pp. 457–468, Feb 2014.

[115] K. Zeng, K. Govindan, and P. Mohapatra, "Non-cryptographic authentication and identification in wireless networks [Security and Privacy in Emerging Wireless Networks]," *Wireless Communications, IEEE*, vol. 17, pp. 56–62, October 2010.

[116] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of Caching Strategies in Modern Cellular Backhaul Networks," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, (New York, NY, USA), pp. 319–332, ACM, 2013.

[117] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *Communications Magazine, IEEE*, vol. 52, pp. 131–139, February 2014.

[118] X. Li, X. Wang, C. Zhu, W. Cai, and V. Leung, "Caching-as-a-Service: Virtual caching framework in the cloud-based mobile networks," in *Computer Communications Workshops (INFOCOM WKSHPS), 2015 IEEE Conference on*, pp. 372–377, April 2015.

[119] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware Virtual Network Function placement for Virtual 5G Network Infrastructure," in *Communications (ICC), 2015 IEEE International Conference on*, pp. 3879–3884, June 2015.

[120] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE Mobile Core Gateways, the Functions Placement Problem," in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, & Challenges*, AllThingsCellular '14, (New York, NY, USA), pp. 33–38, ACM, 2014.

[121] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, "vspace: Vnf simultaneous placement, admission control and embedding," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2018.

[122] H. Guan, T. Kolding, and P. Merz, "Discovery of cloud-RAN," in *Cloud-RAN Workshop*, vol. 2010, 2010.

[123] 3GPP, "Network Sharing; Architecture and Functional Description." TR 23.251.

[124] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *Communications Magazine, IEEE*, vol. 51, pp. 27–35, July 2013.

[125] A. Khan, W. Kellerer, K. Kozu, and M. Yabusaki, "Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence," *Communications Magazine, IEEE*, vol. 49, pp. 134–142, Oct 2011.

[126] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE Mobile Network Virtualization," *Mob. Netw. Appl.*, vol. 16, pp. 424–432, Aug. 2011.

[127] 3GPP, "System Architecture Working Group 1 (SA1) RAN Sharing Enhancements Study Item." TR 22.852.

[128] "Cloud RAN - The Benefits of Virtualization, Centralization and Coordination." White Paper, September 2015.

[129] M. Rahman, C. Despins, and S. Affes, "Configuration cost vs. QoS trade-off analysis and optimization of SDR access virtualization schemes," in *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pp. 1–6, April 2015.

[130] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE wireless virtualization and spectrum management," in *Wireless and Mobile Networking Conference (WMNC), 2010 Third Joint IFIP*, pp. 1–6, Oct 2010.

[131] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN Virtualization in Multi-Tenant LTE-A Heterogeneous Networks," *Vehicular Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[132] G. Tseliou, F. Adelantado, and C. Verikoukis, "Resources negotiation for network virtualization in LTE-A networks," in *Communications (ICC), 2014 IEEE International Conference on*, pp. 3142–3147, June 2014.

[133] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," *Networking, IEEE/ACM Transactions on*, vol. 20, pp. 1333–1346, Oct 2012.

[134] S. Sun, M. Kadoch, L. Gong, and B. Rong, "Integrating network function virtualization with SDR and SDN for 4G/5G networks," *Network, IEEE*, vol. 29, pp. 54–59, May 2015.

[135] M. Peng, S. Yan, and H. Poor, "Ergodic Capacity Analysis of Remote Radio Head Associations in Cloud Radio Access Networks," *Wireless Communications Letters, IEEE*, vol. 3, pp. 365–368, Aug 2014.

[136] S. Khatibi and L. M. Correia, "A model for virtual radio resource management in virtual rans," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 1–12, 2015.

[137] J. Tang, W. P. Tay, and T. Quek, "Cross-layer resource allocation in cloud radio access network," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pp. 158–162, Dec 2014.

[138] N. Nikaein, R. Knopp, L. Gauthier, E. Schiller, T. Braun, D. Pichon, C. Bonnet, F. Kaltenberger, and D. Nussbaum, "Demo: Closer to Cloud-RAN: RAN as a Service," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 193–195, ACM, 2015.

[139] N. Nikaein, R. Knopp, F. Kaltenberger, L. Gauthier, C. Bonnet, D. Nussbaum, and R. Ghaddab, "Demo: OpenAirInterface: an open LTE network in a PC," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, pp. 305–308, ACM, 2014.

[140] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, HotSDN '13, (New York, NY, USA), pp. 25–30, ACM, 2013.

[141] R. Riggio, K. Gomez, L. Goratti, R. Fedrizzi, and T. Rasheed, "V-Cell: Going beyond the cell abstraction in 5G mobile networks," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pp. 1–5, May 2014.

[142] F. Granelli, A. Gebremariam, M. Usman, F. Cugini, V. Stamati, M. Alitska, and P. Chatzimisios, "Software defined and virtualized wireless access in future wireless networks: scenarios and standards," *Communications Magazine, IEEE*, vol. 53, pp. 26–34, June 2015.

[143] A. Dawson, M. Marina, and F. Garcia, "On the Benefits of RAN Virtualisation in C-RAN Based Mobile Networks," in *Software Defined Networks (EWSDN), 2014 Third European Workshop on*, pp. 103–108, Sept 2014.

[144] M. Arslan, K. Sundaresan, and S. Rangarajan, "Software-defined networking in cellular radio access networks: potential and challenges," *Communications Magazine, IEEE*, vol. 53, pp. 150–156, January 2015.

[145] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "FluidNet: A Flexible Cloud-based Radio Access Network for Small Cells," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, MobiCom '13, (New York, NY, USA), pp. 99–110, ACM, 2013.

[146] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating While Computing: Distributed mobile cloud computing over 5G heterogeneous networks," *Signal Processing Magazine, IEEE*, vol. 31, pp. 45–55, Nov 2014.

[147] A. Aissioui, A. Ksentini, A. Gueroui, and T. Taleb, "Toward Elastic Distributed SDN/NFV Controller for 5G Mobile Cloud Management Systems," *Access, IEEE*, vol. 3, pp. 2055–2064, 2015.

[148] M. Usman, A. Gebremariam, U. Raza, and F. Granelli, "A Software-Defined Device-to-Device Communication Architecture for Public Safety Applications in 5G Networks," *Access, IEEE*, vol. 3, pp. 1649–1654, 2015.

[149] Z. Xiao, H. Wen, A. Markham, and N. Trigoni, "Lightweight map matching for indoor localisation using conditional random fields," in *Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on*, pp. 131–142, April 2014.

[150] P. Iovanna and F. Ubaldi, "SDN solutions for 5G transport networks," in *Photonics in Switching (PS), 2015 International Conference on*, pp. 297–299, Sept 2015.

[151] M. Liyanage, I. Ahmad, M. Ylianttila, A. Gurtov, A. B. Abro, and E. M. de Oca, "Leveraging LTE security with SDN and NFV," in *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, pp. 220–225, Dec 2015.

[152] I. Ahmad, S. Namal, M. Ylianttila, and A. Gurtov, "Security in Software Defined Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 2317–2346, Fourthquarter 2015.

[153] W. Ding, W. Qi, J. Wang, and B. Chen, "OpenSCaaS: an open service chain as a service platform toward the integration of SDN and NFV," *IEEE Network*, vol. 29, pp. 30–35, May 2015.

[154] Y. Li, F. Zheng, M. Chen, and D. Jin, "A unified control and optimization framework for dynamical service chaining in software-defined nfv system," *IEEE Wireless Communications*, vol. 22, pp. 15–23, December 2015.