A Cloud-Enabled Small Cell Architecture in 5G Networks for Broadcast/Multicast Services

Michail-Alexandros Kourtis⬤, Bego Blanco, Jordi Pérez-Romero⬤, Dimitris Makris, Michael J. McGrath, George Xilouris, Daniele Munaretto, Rubén Solozabal⬤, Aitor Sanchoyerto⬤, Ioannis Giannoulakis, Emmanouil Kafetzakis⬤, Vincenzo Riccobene, Elisa Jimeno, Anastasios Kourtis, Ramon Ferrús⬤, Fidel Liberal, Harilaos Koumaras, Alexandros Kostopoulos, and Ioannis Chochliouros

Abstract—The evolution of 5G suggests that communication networks become sufficiently flexible to handle a wide variety of network services from various domains. The virtualization of small cells as envisaged by 5G, allows enhanced mobile edge computing capabilities, thus enabling network service deployment and management near the end user. This paper presents a cloud-enabled small cell architecture for 5G networks developed within the 5G-ESSENCE project. This paper also presents the conformity of the proposed architecture to the evolving 5G radio resource management architecture. Furthermore, it examines the inclusion of an edge enabler to support a variety of virtual network functions in 5G networks. Next, the improvement of specific key performance indicators in a public safety use case is evaluated. Finally, the performance of a 5G enabled evolved multimedia broadcast multicast services service is evaluated.

Index Terms—Network function virtualization, software defined networking, edge cloud computing, 5G, MCPTT, eMBMS.

I. INTRODUCTION

THE ADVENT of mobile data networks and various multimedia services has created the need for novel network environments in support of content delivery from an operator's perspective. The continuous evolution and

growth of mobile network data traffic implies a trade-off between increased capacity and cost reduction for the provider (i.e., CAPEX, OPEX). An initial approach to addressing this challenge was investigated by the EU H2020 funded SESAME project [1]. A solution based on a flexible Radio Access Network (RAN) enhanced with Network Function Virtualization (NFV) capabilities was developed by the project. The joint radio-cloud architecture realized the concept of placing intelligence at the network edge, using NFV as an enabler, to build a cost effective and energy-efficient RAN [2]. The ongoing EU funded 5G-ESSENCE project [3] continues to evolve the small cell concept, by integrating compute capabilities (i.e., a low-cost micro servers) at the edge and as a result is able to execute applications and network services, in accordance to the Mobile Edge Computing (MEC) paradigm [4]–[6]. Network Function Virtualization (NFV) [7], [8] serves as a key enabler to this paradigm, as it can provide flexible methods on network service management.

This paper centers on an evolved Cloud Enabled Small Cell (CESC) architecture as proposed in 5G-ESSENCE, whose aim is to enhance the processing capabilities for data that have immediate value beyond locality. Each CESC is comprised of a Small Cell enhanced with a low cost computation resource. The challenges of managing processing-intensive management functions, such as Radio Resource Management (RRM)/ Self Organizing Network (SON) on small cells is also addressed by the project. Also, the project is delivering real-world demonstrations of the solution based on a number of relevant use cases, based on the paradigms of broadcast communications and media delivery.

Under-utilized virtualized resources in small cells will be exploited to their full potential and in a dynamic way [9], in order to support ultra low-latency, high-performance services. The approach being investigated is also expected to improve network resiliency, and to provide substantial capacity gains at the access network for 5G related applications. To achieve these goals, a distributed edge cloud environment (designated as 'Edge Data Centre' -Edge DC-) is under development [10], [11], based on a two-tier architecture. The first tier, i.e., the Light DC, will be distributed among the CESCs, in order to provide latency-sensitive services to users directly from the network's edge. The second tier will be a more centralized, 'high-scale' cloud, namely the Main Data Centre (Main DC), which will provide higher processing

capabilities for computing intensive network applications. It will also have a more centralized view so as to host efficient Quality of Service (QoS) enabled by scheduling algorithms. Both cloud tiers will form the Edge DC in the context of 5G-ESSENCE, which will be viewed as an integrated cloud infrastructure from the perspective of the upper management and orchestration layers.

At the network's edge, each CESC is able to host a number of VNFs comprising a service, which is available to users of a specific operator. Similarly, VNFs can be instantiated inside the Main DC as part of a Service Function Chaining (SFC). The Light DC can be used to implement different functional splits within the Small Cells as well as supporting end user mobile edge applications. At the same time, 5G-ESSENCE proposes the development of small cell management functions implemented as VNFs, which run in the Main DC and coordinate a fixed pool of shared radio resources, instead of considering each small cell as a set of unique and dedicated resources.

It should be noted that this paper does not only propose the development and adaptation of a multitenant CESC platform. It also addresses several current aspects that affect performance in 5G virtualized environments such as:

- Conformity of the proposed architecture to the Radio Resource Management (RRM) in the evolving Next Generation RAN (NG-RAN) architecture.
- Local break-out to support a plethora of VNFs at the edge.
- A public safety use case and evaluation of the improvement of specific KPIs.
- A broadcast media delivery use case, evaluating the eMBMS technology deployed at the edge.

Last but not least, it is worth noting that the two-tier architecture proposed in this manuscript is well aligned with the current views on 5G architecture as described by the 5G-PPP in [13], where infrastructure programmability and the split between the control and user planes are identified as two key logical architecture design paradigms for 5G.

The remainder of the paper is organized as follows. Section II describes the overall 5G-ESSENCE architecture and focuses on the two tier virtualized execution environment. It also proposes an integrated management solution and finally, it specifies the tier, in which a VNF should run, in order to achieve better performance. Section III describes how the proposed architecture matches the evolving 5G New Radio (NR), taking into account its basic components of gNB-DU and gNB-CU. Section IV describes a solution that has been developed, which allows any type of VNF to be placed at the edge of the network, addressing the local break-out issue. Section V examines how the proposed architecture is deployed in a public safety use case and more specifically it evaluates the deployment of a Mission Critical Push to Talk (MCPTT) application over 5G infrastructures, using specific KPIs. Section VI demonstrates and evaluates the benefits of a broadcast media delivery service when deployed at the edge, by exploiting the key features of the eMBMS technology. Finally, Section VII concludes the paper and draws future research lines.

II. 5G-ESSENCE OVERVIEW

This section, firstly, details the overall 5G-ESSENCE architecture, and then, reviews the definition of a network service in the context of the project.

A. 5G-Essence Overall Architecture

In the 5G ESSENCE approach, the Small Cell concept is evolved not only to provide multi-operator radio access [14], but also, to achieve an increase in the capacity and the performance of current RAN infrastructures, and to extend the range of the services provided while maintaining its agility. To achieve these ambitious goals, 5G ESSENCE leverages the paradigms of RAN scheduling and additionally provides an enhanced, edge-based, virtualized execution environment attached to the small cell, taking advantage and reinforcing the concepts of MEC and network slicing.

The 5G-ESSENCE architecture is shown in Figure 1. It combines the current 3GPP framework for network management in RAN sharing scenarios and the ETSI NFV framework for managing virtualized network functions [15]. The CESC offers virtualized computing, storage and radio resources and the CESC cluster is considered as a cloud by the upper layers. This cloud can also be 'sliced' to enable multi-tenancy. VNFs which implement the different Small Cells features, as well as supporting end user mobile edge applications are supported by the execution platform.

As shown in Figure 1, the 5G-ESSENCE architecture allows multiple network operators (tenants) to provide services to their users through a set of CESCs deployed, owned and managed by a third party (i.e., the CESC provider). In this way, operators can extend the capacity of their 5G RAN in areas where the deployment of their own infrastructure could be expensive and/or inefficient, as would be the case (e.g., highly dense metropolitan areas) where massive numbers of Small Cells would be required to provide expected high quality services within a given geographical area.

In addition to capacity extension, the 5G-ESSENCE platform is equipped with a two-tier virtualized execution environment, materialized in the form of an Edge DC, which supports the provisioning of MEC capabilities for mobile operators [16]–[18], in order to provide an enhanced user experience and agile service delivery. The first tier, i.e., the Light DC hosted inside the CESCs, is used to support the execution of VNFs for carrying out the virtualization of the Small Cell access. In this regard, network functions supporting traffic interception, GTP encapsulation/ decapsulation and some distributed RRM/SON functionalities are expected to be executed therein. VNFs that require low processing power, e.g., a Deep Packet Inspection (DPI), a Machine-to-Machine (M2M) Gateway, and so on, could also be hosted here. The connection between the Small Cell Physical Network Functions (PNFs) and the Small Cell VNFs can be realized through, e.g., the network Functional Application Platform Interface (nFAPI). Finally, backhaul and fronthaul transmission resources are part of the CESC, allowing for the required connectivity.

The second cloud tier, i.e., the Main DC, will host more computation intensive tasks and processes that require centralization in order to have a global view of the underlying infrastructure. This includes the cloud-enabled Software Defined – Radio Access Network (cSD-RAN), controller which provides the control plane decisions for all the radio elements within the geographical area of a CESC cluster, including the centralized Radio Resource Management (cRRM) over the entire

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

KOURTIS *et al.*: CESC ARCHITECTURE IN 5G NETWORKS FOR BROADCAST/MULTICAST SERVICES 3
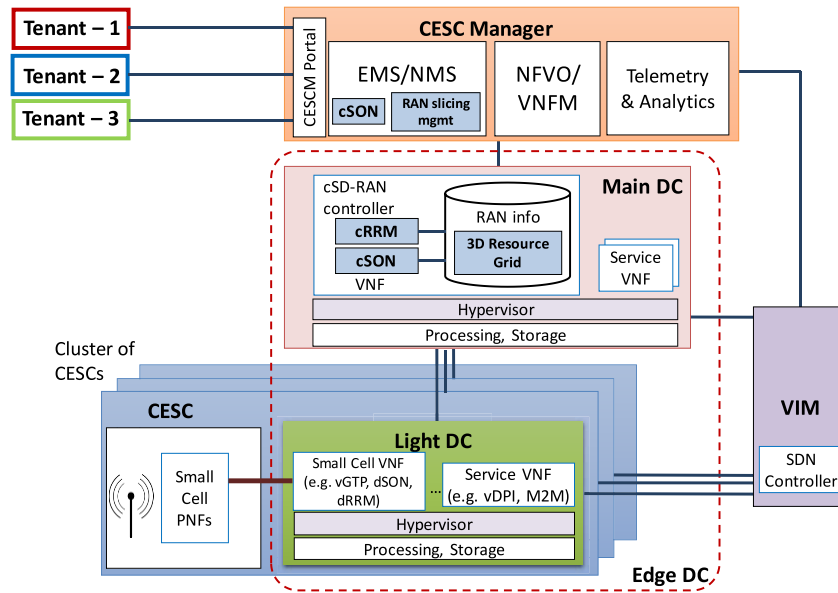

Fig. 1. 5G ESSENCE High Level Architecture.

CESC cluster, which is presented in detail in Section V. Other potential VNFs that could be hosted by the Main DC include security applications, traffic engineering, mobility management, and in general, any additional network E2E services that can be deployed and managed on the 5G-ESSENCE virtual networks, effectively and on demand.

The management modules for the operation of the CESC platform and service provisioning within the CESCM framework are shown in Figure 1. The following subsections provide a more detailed description of each architecture component in Figure 1.

B. Main Architectural Components of 5G ESSENCE

In our scope, a CESC consists of a Multi-RAT 5G small cell with its standard backhaul interface, standard management connection (TR069 interface for remote management) and with the necessary modifications to the data model (TR196 data model) to allow Multi-Operator Core Network (MOCN) radio resource sharing. The CESC is composed by a physical small cell unit attached to an execution platform based on x86 processors. Edge cloud computing and networking are realized through sharing the computation, storage and network resources of the micro servers present in each CESC and from the Light DC Therefore, the CESCs become a neutral host for network operators or virtual network operators that want to share IT and network resources at the edge of the mobile network.

The CESC is envisioned to accommodate multiple operators (tenants) by design, offering Platform as a Service (PaaS), capable of providing the deployed physical infrastructure among multiple network operators. Different VNFs can be hosted in the CESC environment for different tenants. This also provides support for mobile edge computing applications deployed for each tenant that, operating in proximity to the end users, may significantly reduce the service delivery time and deliver composite services in an automated manner.

Moreover, the CESC is the termination point of the GTP-User Plane (GTP-U) tunneling which encapsulates user IP packets from the core network entities (e.g., the Evolved Packet Core (EPC) Serving Gateway (SGW) in LTE) destined to the User Equipment (UE) and vice versa.

The CESC exposes different views of the network resources: per-tenant small cell view, and physical small cell substrate, which is managed by the network operator, decoupling the management of the virtual small cells from the platform itself. The CESC provides termination of multiple S1 interfaces connecting the CESC to multiple MME/SGW entities as in S1-Flex. The interconnection of multiple CESCs forms a 'cluster' which can facilitate access to a broader geographical area with one or more operators (even virtual ones), extending the range of their service footprint, while maintaining the required agility for on demand extensions.

1) The Edge DC Encompassing Main DC and Light DC: The proposed architecture is combining the concepts of MEC and NFV with Small Cell virtualization in 5G networks and enhancing them for multi-tenancy support. The purpose of the Edge DC will be to provide Cloud services within the network infrastructure and also to facilitate, by promoting and assisting, the exploitation of network resource information. To this end, all the hardware modules of the Light DC and the Main DC will be delivered as abstracted resources using novel virtualization techniques. Both networking and computing virtualization extensions will be developed using open frameworks such as OPNFV. The combination of the proposed Edge DC architecture coupled with the concepts of NFV and SDN will facilitate greater levels of flexibility and scalability.

As seen in the architecture presented in Figure 1, the Main DC will be able to execute different Small Cell and Service VNFs under the control of the CESCM. In particular, the Main DC hosts the cSD-RAN controller which performs cRRM decisions for handling efficiently the heterogeneous access network environment composed of different access technologies such as 5G RAN, LTE, and Wi-Fi. These radio

access networks can be programmable and under the supervision of the centralized controller. The cSD-RAN controller updates and maintains the global network state in the form of a database called "RAN Information", which includes, among other elements, an abstraction of the available radio resources in the CESC cluster. This abstraction takes the form of a "3D Resource Grid" that characterizes the resources in the domains of time/space/frequency. The RAN information will be used by the cRRM to perform resource allocation decisions such as scheduling.

The cSD-RAN controller can also host centralized Self Organizing Network (cSON) functionalities that are required for coordinating multiple small cells. These types of functionalities are not appropriate to run in the Light DC, as for example, InterCell Interference Coordination (ICIC) functions. Other distributed (dSON) functions and/or distributed RRM (dRRM) functions that are of low complexity and that do not involve the coordination of multiple small cells will run at the Light DC. For example, this could be the case of an admission control function that only takes decisions based on the current load existing at a given cell.

2) The CESCM: The CESC Manager (CESCM) is responsible for coordinating and supervising the consumption, performance, and delivery of radio resources and services. It controls the interactions between the infrastructure elements (CESCs, Edge DC) and network operators. It also has responsibility for Service Level Agreements (SLAs) compliance. From an architectural perspective, the CESCM encompasses telemetry and analytics as fundamental capabilities to enable efficient management of the overall infrastructure landscape and network. The Virtualized Infrastructure Manager (VIM) is responsible for controlling the NFV Infrastructure (NFVI), which includes the computing, storage and network resources of the Edge DC.

Management and orchestration of the proposed uniform virtualized environment, which is able to support both radio connectivity and edge services, is a challenging task by itself. Management of the diverse lightweight virtual resources is of primary importance, in order to enable a converged cloud-radio environment and efficient placement of services, and extend current solutions in the NFV field [19]–[21]. For this purpose, the CESCM is the central service management and orchestration component within the architecture. Generally speaking, it integrates collectively the traditional network management elements, and the new functional blocks necessary to realize NFV operations. A single instance of CESCM is able to operate over several CESC clusters at different Points of Presence (PoP), each constituting an Edge DC through the use of a dedicated VIM per cluster.

An essential component of the CESCM is the Network Functions Virtualization Orchestrator (NFVO). It is responsible for realizing network services on the virtualized infrastructure and includes the interfaces to interact with the CESC provider for service management (e.g., exchange of network service descriptors). The NFVO composes service chains (comprising of two or more virtualized Networks Services located in one or more CESCs that jointly realize the complete orchestration service) and manages the deployment of VNFs at the Edge DC. The NFVO uses the services exposed

by the VNF Manager (VNFM), which are in charge of the instantiation, updating, querying, configuring and termination of VNFs. Moreover, NFVO may include features to enhance the overall system performance, e.g., to improve energy efficiency.

The CESCM hosts also the Element Management System (EMS), which provides a package of end-user functions for the management of both physical network functions (PNFs) and VNFs at the CESCs. In particular, the EMS carries out key management functionalities such as Fault, Configuration, Accounting, Performance, Security (FCAPS) operations. The EMS has responsibility for partitioning the single whole-cell management view into multiple virtual-cell management views, one per tenant. In this way, a virtualized Small Cell with a set of (limited) management functionalities can be made visible to, e.g., the Network Management System (NMS) of each tenant, in order to collect performance counters, configure neighbor lists for a proper mobility management, etc. In addition to the NMSs of each tenant the CESCM can also incorporate an NMS for managing the whole set of CESCs deployed by an operator. This may be appropriate for example in scenarios where existing CESCs belonging to different vendors in the same deployment, each one with its own EMS. The EMS/NMS will also host the cSON functionalities (e.g., self-planning, Coverage and Capacity Optimization (CCO), etc.) and the functionalities for the life-cycle management of RAN slicing (i.e., for the creation, modification or termination of RAN slices).

As shown in Figure 1, the CESCM encompasses a telemetry and analytics module that capture and analyzes relevant indicators of infrastructure and network operations. This capability provides the CESCM with accurate operational knowledge that characterizes the behavior of the network and its users in relation to the utilization of both cloud and radio resources. This will facilitate the realization of effective optimization approaches based on, e.g., machine learning techniques for service placement, which can dynamically adapt to the context of the provided services and their execution environment and to enable automated enforcement of SLAs. Finally, the CESCM also incorporates the CESCM portal. It is a control panel with Web Graphical User Interface (GUI) that serves as the entry point for all users, including the CESC provider and tenants, to CESCM functionalities and constitutes the main graphical frontend to access the 5G ESSENCE platform. The CESCM Portal in general provides visualization of the monitoring information of the platform, the agreed SLAs, and the available network services/VNFs, allowing parameters' configuration.

3) The VIM: The CESCM functions will be built upon the services provided by the VIM which has responsibility for appropriately managing, monitoring and optimizing the overall operation of NFVI resources (i.e., computing, storage and network resources) at the Edge DC. The role of VIM is essential for the deployment of NFV services and to form and provide a layer of NFV resources for CESCM functions. NFVI resources will be ultimately offered as a set of APIs that will allow the execution of network services over the decentralized CESCs, located at the edge of the network. As seen in Figure 1, the VIM relies on an SDN

controller for interconnecting the VNFs and for offering SFC on the data-plane by establishing the path for the physical connections.

III. RADIO RESOURCE MANAGEMENT FUNCTIONS IN 5G

The operation of the radio interface in the RAN is controlled by means of Radio Resource Management (RRM) functions that regulate the assignment of radio resources to the existing radio bearers, which are the data transfer services delivered by the radio protocol stack, in accordance with their expected level of Quality of Service (QoS). Main RRM functions include i) Radio Admission Control (RAC), which decides on the acceptance/rejection of new radio bearers, ii) Radio Bearer Control (RBC), which controls the establishment, maintenance and release of radio bearers and the associated configuration of radio resources, iii) Connection Mobility Control (CMC), which deals with cell reselection for users in idle mode and with handover for users in connected mode, and iv) Packet Scheduling (PS), which performs the dynamic assignment of Physical Resource Blocks (PRBs) to the different radio bearers to perform transmissions through the radio interface. This section describes different options for placing the RRM functions in the proposed architecture and considering the relationship between this architecture and the Next Generation - Radio Access Network (NG-RAN) architecture of the 5G system [22].

In the current NG-RAN architecture, RRM functionalities are hosted by the gNB nodes, which provide the 5G New Radio (NR) user and control plane terminations. The gNB can be decomposed in a gNB Central Unit (gNB-CU) and one or more gNB Distributed Units (gNB-DU). Both units are interconnected through the F1 interface. The operation of the gNB-DU is partly controlled by the gNB-CU. One cell is supported by only one gNB-DU, but one gNB-DU can support multiple cells. According to the current release 15 of 3GPP specifications, the functional split between the gNB-CU and the gNB-DU considers that the gNB-CU hosts the Radio Resource Control (RRC), Service Data Adaptation Protocol (SDAP) and Packet Data Convergence Protocol (PDCP) layers while the gNB-DU hosts the Radio Link Control (RLC), Medium Access Control (MAC) and Physical (PHY) layers. This corresponds to the so-called split option 2 [23]. However, it is envisaged that future releases of the specifications will incorporate other lower layer splits (e.g., MAC-PHY and intra-PHY split) that are currently under study in [24].

The proposed architecture described previously in Section II (see Figure 1) enables a distributed deployment of the RAN functionality across the dedicated hardware embedded in the CESC (i.e., small cell PNF) and the distributed virtualized execution environment provided by the Edge DC with lightweight NFVI Points of Presence (PoP) embedded in the CESC units (e.g., Light DC) and an external centralized NFVI PoP (i.e., Main DC). Such infrastructure configuration can accommodate different deployment options such as the gNB decomposition between a gNB-CU and one or more gNB-DUs considered in the 3GPP NG-RAN. In general terms, it can be envisaged that the main DC can host the functionalities of the gNB-CU
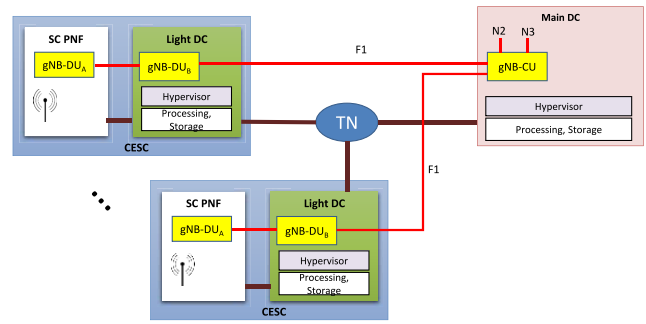

Fig. 2. Example 1 for mapping the gNB-CU and gNB-DU entities in the proposed architecture.
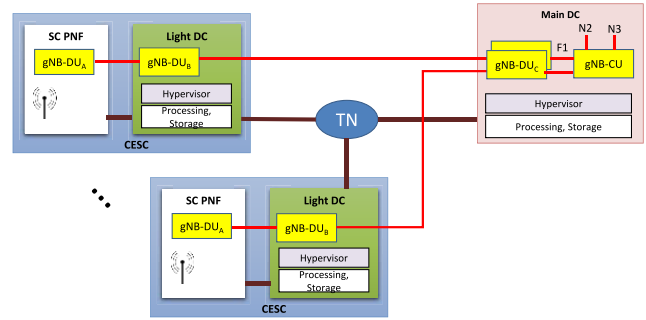

Fig. 3. Example 2 for mapping the gNB-CU and gNB-DU entities in the proposed architecture.

in addition to some other functionalities of the gNB-DU with higher computational requirements.

To illustrate the flexibility of the proposed architecture to support different splits between gNB-CU and gNB-DU some examples are given in Figure 2 to Figure 4. For simplicity, in all these figures only the CESC and the main DC components of the 5G ESSENCE architecture are shown, together with the Transport Network (TN) used to interconnect the different CESCs and the main DC. The different logical interfaces between entities are depicted in red.

The example of Figure 2 considers the 3GPP functional split option 2 in which the gNB-CU runs at the main DC and incorporates the SDAP, PDCP and RRC protocols (thus hosting RAC, RBC and CMC functionalities), while the gNB-DU runs at the CESCs. In this case, taking advantage of the light DC, the gNB-DU is further split into two parts, the gNB-DU$_A$, which executes the physical layer and is implemented at the SC PNF, and gNB-DU$_B$, which executes the MAC and RLC layers and runs at the local microservers of each CESC composing the light DC. In this example, the PS functionality, which runs at the MAC layer, will be distributed at each CESC. In contrast, the example of Figure 3 assumes also the functional split option 2 but considering that the MAC functionality is further split between the gNB-DU$_B$ running at the light DC and hosting the lower level MAC, e.g., Hybrid Automatic Repeat reQuest (HARQ) process, and the gNB-DU$_C$, running at the main DC and hosting the PS and RLC level. With this example, a coordinated scheduling for multiple cells can be implemented at the main DC. Therefore, this approach would facilitate the implementation of techniques such as Coordinated Multi-Point (CoMP) and will facilitate
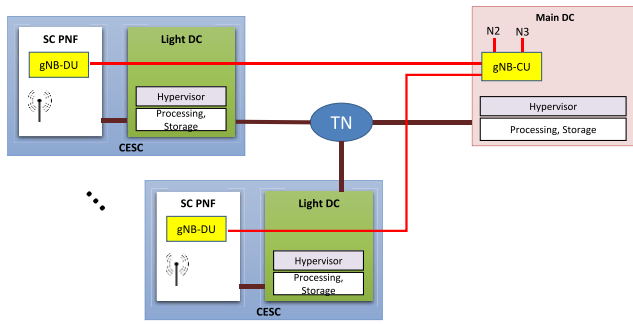
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6   IEEE TRANSACTIONS ON BROADCASTING

Fig. 4. Example 3 for mapping the gNB-CU and gNB-DU entities in the proposed architecture.

Fig. 5. Overview of the proposed Edge Enabler solution for "local-breakout".

InterCell Interference Coordination (ICIC). In addition to the gNB-DU$_C$ function, the Main DC also hosts the gNB-CU with the RAC, RBC and CMC functionalities that can also be coordinated for multiple cells.

As another example, Figure 4 assumes the MAC-PHY functional split. In this case, the gNB-DU includes the physical layer and is executed at the CESC PNF, while the gNB-CU includes the MAC layers and above and is executed at the main DC. Like in the previous example, also in this case the PS function can have a coordinated vision of multiple cells. It is worth mentioning that, in this example, the interface between the gNB-DU and the gNB-CU would be the one defined by 3GPP when specifying the lower layer split, since current interface F1 is just for the split option 2.

Summarizing all presented mapping options and based on the current activities of various 5G enablers, the current preferred method is the third option, as it does not limit the processing requirements of the 5G New Radio (NR). The other options are still in early stages of development but will allow future 5G-enabled systems to further extend their capabilities. The split of the Small Cell's PF functions and their distribution across various hosting nodes, will also ensure resiliency and scalability depending on the requirements of each deployment scenario.

IV. EDGE ENABLER

In the scope of this paper, the issue of network service management at the edge is also discussed. Mobile Operators are facing an ever growing volume of video in their networks and demanding requirements from their most sophisticated customers for 5G low latency services. Mobile edge applications have been the long promised solution, however, such deployment has always been a challenging task as previous solutions [25], [26] did not provide a simple mechanism for selectively off-loading traffic. One of the main problems identified in the convergence of NFV and 5G is the General Packet Radio Service (GPRS) Tunneling Protocol (GTP) packet handling. GTP packets are used to allow end user's mobility. In the communication between the EPC and the SC, the data packets exchanged are GTP encapsulated. When these packets are forwarded towards an NFV environment, the VNFs will fail to handle them appropriately. In this scenario, the packets need to be decapsulated from their outer GTP header forwarded to the VNF and then if needed be re-encapsulated and sent to the UE.
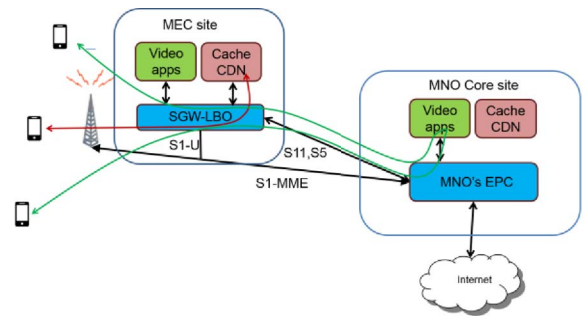
In this work this requirement is addressed by deploying a new Serving Gateway (S-GW) that enables traffic local break-out (namely, SGW-LBO [27]). The SGW-LBO is a modified SGW core function which has been enhanced to allow traffic to be "broken-out" and steered locally, whilst maintaining 3GPP compatibility, to support i) caching of various popular applications and other content and ii) other applications that require low latency or local offload (e.g., smart cities or autonomous cars). From the perspective of a Mobile Operator, the solution is transparent to and does not require modifications in the radio or core network. It can also be software upgraded to create a future-proof bridge to 5G. Figure 5 shows an overview of the solution implemented to deal with the GTP packet handling at the edge. Furthermore, the creation of duplicate instances of S-GW and their placement at the edge, results in a distributed EPC solution, which also solves the packet handling at the VNF level, as the packets that arrive will be GTP decapsulated.

This fully NFV-based solution, deployed at the edge cloud on Commercial Off-the Shelf (COTS) servers and virtualized infrastructure, is easy to deploy and manage, reduces costs, and enables new services and revenue models, which are critical to the economics of edge deployments.

This MEC solution acts as data plane routing function and enables local breakout based on per-user or per-traffic stream policies provisioned via APIs.

The solution, located at the edge, connects to the mobile network operator's MME and PGW via 3GPP standard interfaces, S11 and S5 respectively. The SGW-LBO is a standard compliant 3GPP SGW node, part of the MEC platform that is controlled and coordinated by the operator from the central core.

This solution is key to offer 5G-ready performance and use cases with LTE and NSA 5G-NR radio, while building the network edge as a crucial stepping stone to the 5G.

V. PUBLIC SAFETY USE CASE: DEPLOYMENT OF AN MCPTT SERVICE ON 5G ESSENCE ARCHITECTURE

This section introduces the deployment possibilities of a Mission Critical Push To Talk (MCPTT) service taking advantage of the 5G ESSENCE architecture. A Push-to-talk (PTT) functionality allows half-duplex instantaneous individual and broadcast communication using a button to switch between reception mode and transmission mode.

By extension, MCPTT points to a PTT solution that meets the strict requirements of public safety mission critical communications, including high availability, reliability, low latency and other QoS indicators. Both NFV and MEC capabilities of the proposed platform provide the appropriate environment to guarantee the strict KPIs of mission critical communications.

In order to support the demanding set of requirements of the Public Safety (PS) use-case, 5G ESSENCE project proposes a Software Defined Radio Access Network (SD-RAN) architectural framework that leverages NFV and SDN to provide a complete virtualized ecosystem suitable for the execution of Virtual Network Functions (VNF). This way, building 5G–based Mission Critical (MC) communication services at the Edge DC makes it possible to relieve the core network from the traffic that can be efficiently processed and served closer to the end user and, thus, to reduce the response time, which favors the service itself as well. To this aim, the decoupling of the control and user planes of the different stacks frees from the enormous fronthaul latency restrictions. Additionally, the use of end-to-end network slicing mechanisms will allow sharing the infrastructure among other vertical industries or services and customizing its capabilities on a per-tenant basis.

Mission-Critical Push-to-Talk (MCPTT) service, as a relevant component of MC communications, is an IP-based MC service that can benefit from the proposed architecture. This kind of service requires of a Session Initiation Protocol (SIP) core such as IP Multimedia Subsystem (IMS) to operate. An IMS aims to reach interoperability for session control in all-IP Next Generation Networks and is currently implemented as a centralized subsystem attached to the Evolved Packet Core (EPC) of each operator.

Adopting the proposed architecture, MC services can be distributed near the end-user, making use of MEC capabilities obtained throughout the virtualization of the resources in the Edge DC and also with the exploitation of the Edge Enabler service described in Section IV. This distribution of services in the proximities of the end-user benefits first responders from reduction of call setup times, responsiveness of the communications and mouth-to-ear latency. Thanks to NFV, it also enables service providers to scale the user plane on demand in specific locations straightforwardly.

Depending on the characteristics of the target service architecture, there are two alternatives in order to bring a MCPTT service to the edge. Due to the complexity of the MCPTT service and the associated service management functions, the first service deployment scheme considers that most of the administrative and management tasks remain at the core network and only the media-related user plane is transferred to the edge. The second deployment scheme considered brings the whole MCPTT system to the edge, requiring a higher number of edge resources.

Next, we are going to analyze the impact of the proposed architecture on the performance of the MCPTT service described. We will compare the two aforementioned deployments to a conventional centralized MCPTT service over current 4G networks, in order to illustrate the performance improvements that can be achieved by leveraging the proposed architecture when deploying services at the edge of the network.

TABLE I
MCPTT SERVICE KPIS

MCPTT KPIs	Threshold	Likelihood	LTE Packet Delay Budget
KPI 1 MCPTT Access Time	< 300 ms	95% of all MCPTT requests	< 60 ms
KPI 2 E2E MCPTT Access Time	< 1000 ms	N/A	< 60 ms
KPI 3 Mouth-to-Ear Latency	< 300 ms	95% of all voice bursts	< 75 ms
KPI 4 Late Call Entry Time (encrypted calls)	< 350 ms	95% of Late Call entries	< 60 ms

A. MCPTT Service KPIs

As previously introduced, MC communications have stringent performance requirements, especially regarding to call set-up time and mouth-to-ear delay in both unicast and broadcast communications. 3GPP TS 22.179 [28] defines these main KPIs along with the target values, which are gathered in Table I.

The MCPTT Access Time (KPI 1) is defined as the time between when an MCPTT User requests to speak and when this user gets a signal to start speaking. This time does not include confirmations received from users. It should be noted that the MCPTT Access time does not include the time for an MCPTT user to affiliate to the group since we consider the common case where group affiliations are performed beforehand. The End-to-end MCPTT Access Time (KPI 2) is defined as the time between when an MCPTT User requests to speak and when this user gets a signal to start speaking, including acknowledgement from first receiving user before voice can be transmitted. The Mouth-to-Ear Latency (KPI 3) is the time between an utterance by the transmitting user, and the playback of the utterance at the receiving user's speaker. The Late Call Entry Time (KPI 4) refers to the time needed for a registered user to get into an ongoing call.

Each of these values is tied to a likelihood value and to a LTE Packet Delay Budget. The significance of this value is that it determines the maximum acceptable packet of the LTE part, i.e., from the time a packet leaves the UE until it leaves the P-GW in a conventional centralized service. Thus, the purpose of this use case is to demonstrate the further reduction of these KPIs deducing the portion of latency that would be avoided thanks to the deployment of the MCPTT service at the edge of the network. The following sections discuss the obtained results.

B. Conventional Centralized Service Over 4G Networks

To that aim, the proposed architecture infrastructure is used to characterize current delays between the UEs and the traditional MCPTT ASs deployment behind the EPC of the core network. Besides, a server in France has been employed with all the necessary MCPTT ASs.

After the measurement of the current latencies in a wide variety of scenarios, the average call set-up time (i.e., MCPTT Access Time – KPI 1) obtained through a centralized MCPTT server located behind a conventional EPC core is 190 ms (not considering the UE processing time).
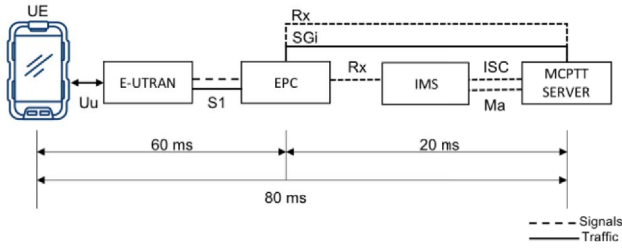
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

8   IEEE TRANSACTIONS ON BROADCASTING

Fig. 6.   Distribution of the delays in a conventional centralized 4G network.

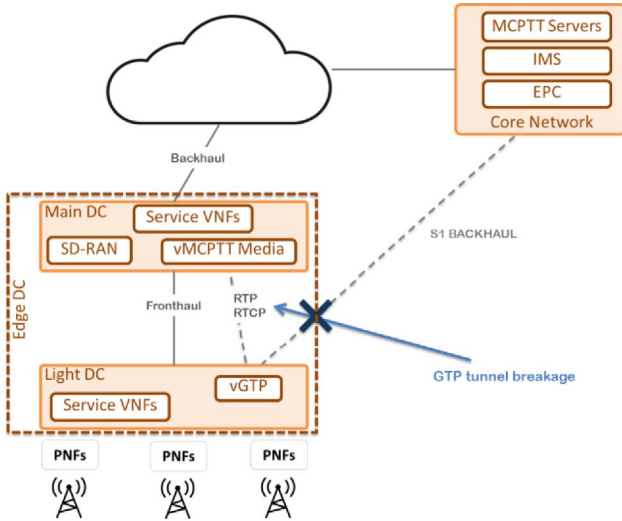Fig. 7.   Distributed MCPTT user plane at the Edge DC.

Fig. 8.   Distributed MCPTT complete service at the Edge DC.

C. Distributed MCPTT User Plane at the Edge DC

Figure 7 shows the first possibility, where there is a centralized MCPTT service in the core network for OAM and call signaling, though distributing MCPTT user plane replications (MCPTT Participating and Controlling Application Servers) along the Edge DC. In this configuration, the pure control elements of the service, such as MCPTT CMS, GMS or IdMS, are exclusively located behind the EPC, but call managers have MCPTT user plane copies of the edge of the network to help scaling the service nearer to the end user, achieving to reduce "mouth-to-ear" latency of the service.

With this scheme, OAM and call signaling procedures are performed against the centralized instance of the service. During the call set-up, the central MCPTT instance will assign the MCPTT Application Server (MCPTT AS) that will manage the call. In this particular case, the assigned server is located in the Edge DC and will manage the calls among the users attached to the cell cluster. From this moment on, the UEs know the destination IP of the MCPTT AS they have to connect to. Instead of forwarding the user data as usual through the GTP tunnel towards the central servers, the RTP/RTCP traffic is redirected to the local MCPTT AS. To this aim, and in order to avoid the breaking of the GTP tunnel, the deployment of an edge enabler that allows selective traffic off-load becomes of utmost interest. In this sense, Section IV introduces an innovative Serving Gateway (S-GW) that enables traffic local break-out.
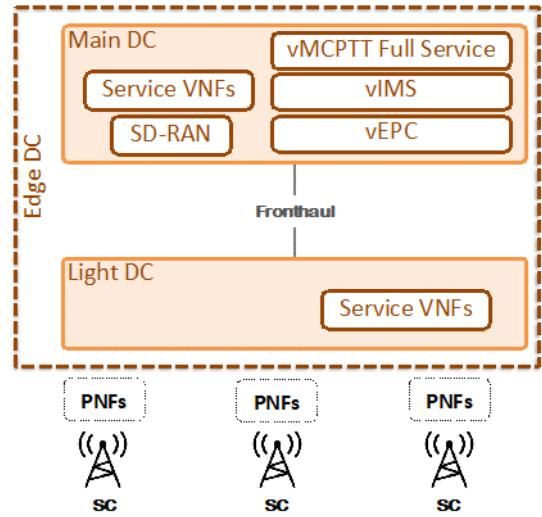
Considering the first deployment option of the MCPTT service, which is shown in Figure 7, it is noted that, in this case, only the user plane has been moved to the edge. Hence, KPI 1 will not be affected since all the signaling required to set up a call would be dealt by the MCPTT servers located behind the EPC in the core network. In contrast, the mouth-to-ear latency (KPI 3) related to the user plane of the service would benefit from proximity to the user location of the vMCPTT Media VNF in the Edge DC. Given the fact that the measured processing delay of the MCPTT VNF user plane is 10 ms, the cSD-RAN controller has approximately 20 ms improvement margin to perform its tasks and the local break-out process.

D. Distributed MCPTT Complete Service at the Edge DC

The second deployment alternative, depicted in Figure 8, corresponds to a tactical bubble. The isolation of this option requires the deployment of the complete MCPTT service locally.

In this case, the use case considers the deployment of a fully distributed service at the edge, without the need to connect to the central service at all. This implies that the local MCPTT service needs to provide not only the media plane support but all the OAM and call signaling features.

If the local users need to communicate with external users, this should be considered as MCPTT communications between different administrative domains, requiring inter-AS communications.

Considering this second deployment option of the service as a standalone and isolated full MCPTT service, in addition to the commented improvement over KPI 3, there is also an improvement possibility over KPI 1. In fact, considering the delay introduced by the control plane of the MCPTT VNF the call set-up time is 85 ms, cSD-RAN controller has about 200 ms improvement margin to complete control plane tasks.

The MCPTT use-case offers an important aspect in the proposed 5G architecture as it covers the area of security and emergency communications in 5G. Furthermore, the presented
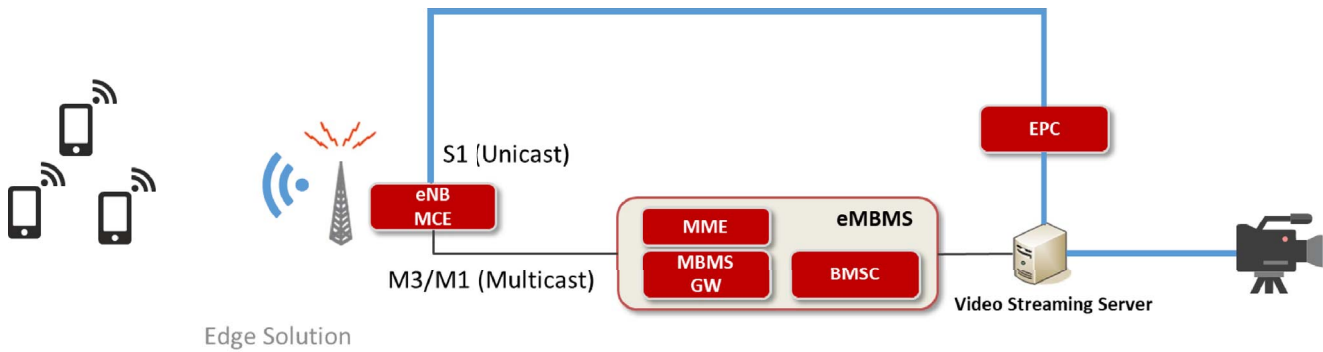
Fig. 9.   Overview of the experimental platform for Unicast and eMBMS video service deployed at the edge.

evaluation results not only show the advantages of a distributed NFV-enabled deployment, but also the importance of the aforementioned Edge solutions.

VI. eMBMS at the Edge

The second use-case presented in this paper is related to media content broadcasting across multiple users. This section presents the architecture of the proposed service, analyses the use of the eMBMS service and evaluates its performance compared to a unicast-based media content delivery service. The core of the proposed architecture is the functionality of an edge solution, as elaborated in 5G ESSENCE project, which allows the provision of LTE unicast and broadcast communications.

With the advent of the LTE technology, eMBMS has become an attractive solution for network operators which want to increase bandwidth capacity and improve service quality while maintaining low-cost investments for upgrading the network architecture. eMBMS is an optimized broadcast/multicast service, which leverages on a point-to-multipoint link to transmit control/data information from the Cloud-enabled Small Cell to a group of users. Related studies have addressed the key role that eMBMS has in the evolving 5G environment [30], [31]. In this paper we study the efficiency and performance of a virtualized eMBMS network service, when deployed at the edge to relieve the core network.

The deployment at the edge can improve the performance of the eMBMS service, as it minimizes delay. The video distribution service is brought next to the users, therefore quality degradation due to network conditions is minimized. In the alternative case where the eMBMS service would be deployed behind the EPC and not close to the eNB, unstable network conditions on the backhaul link and additional delay can gravely affect the service's performance.

The experimental process is based upon the designed system architecture used, which is depicted in Figure 9. For the purposes of this study, and in order to emulate a large crowd live event, a live camera feed was used to deliver live media content to the UEs, over a Cloud-enabled Small Cell infrastructure. The Small Cell and the eMBMS modules are located close to the place of the crowd event. The EPC is located far away, at the premises of the mobile network operator and is connected to the eNodeB through the S1 backhaul link. The video stream was transmitted using 2 modes.

- Unicast: The camera live feed is transcoded in the Video Streaming Server at the appropriate bit rate and then delivered through the EPC to the Small Cell and the UEs. For each user a unique network data flow is created.
- eMBMS (Broadcast): The camera live feed is directly sent from the Video Streaming Server to the eMBMS module (bypassing the EPC), which creates a multicast stream that each UE can watch after it joins the multicast group.

The UEs required for this experiment need to have special capabilities, i.e., to support broadcast functionalities in their hardware and to have installed an appropriate middleware to receive broadcast traffic. These UEs are not easily found in the market. For the needs of this paper, only three UEs were found, from the 5G-ESSENCE project. They were manufactured by Bittium, and also have the Expway middleware installed, in order to be able to receive broadcast traffic from the Small Cell. Furthermore, the CESCs were built based on NOKIA small cells, with the eMBMS feature enabled. Additionally, the EPC was properly configured for the eMBMS scenario, to allow the traffic to go directly from the Video Streaming Server to the eMBMS module, as shown in Figure 9.

The eMBMS service solution is implemented following the 3GPP TS 23.246 [29], and includes the following modules in the mobile core network domain.

- Mobility Management Entity (MME): It supports the session control of eMBMS bearers and the transmission of session control messages towards multiple Multi-cell / Multicast Coordination Entities (MCEs), which are hosted in the radio device;
- MBMS Gateway (MBMS-GW): It is present between the BMSC (see below) and eNBs and is in charge of sending/broadcasting eMBMS packets to each eNB transmitting the service. The MBMS GW uses IP Multicast as the means of forwarding MBMS user data to the eNB. The MBMS GW performs MBMS Session Control Signaling (Session start/stop) towards the E-UTRAN via MME;
- BMSC (Broadcast Multicast Service Center): It ingests content (both live streams from encoders and push files from a content management system) and distributes them to the MBMS-GWs. It controls eMBMS sessions, i.e., initiates session control messages (e.g., Session Start/Stop) towards MBMS-GWs. It provides commands and instructions over the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10   IEEE TRANSACTIONS ON BROADCASTING
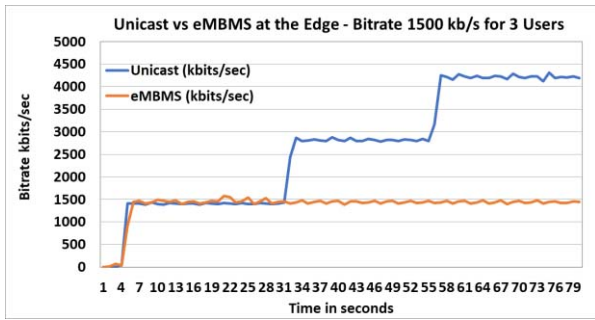

Fig. 10. Performance evaluation for Unicast vs eMBMS at the edge, in terms of bit rate of S1 link, in a 1500kb/s bitrate scenario, with 3 simultaneous users.
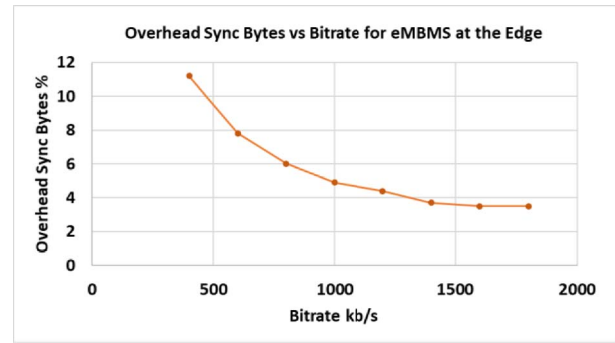

Fig. 11. Variation of the percentage of overhead Sync Bytes vs video stream bit rate, in eMBMS at the edge scenario.

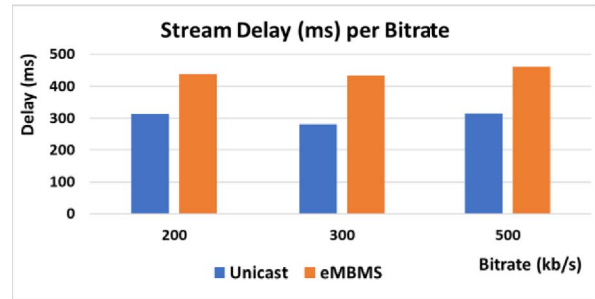Fig. 12. Comparison of content delivery latency for Unicast and eMBMS modes, for various encoding bit rates.

control plane, so that the MBMS gateway can send content to the appropriate eNBs in the selected areas.

The eMBMS service provides broadcasters the ability to stream content to a large group of users, while maintaining a constant required bandwidth, regardless of the number of users.

A. eMBMS Network Service Evaluation

As the primary goal of an eMBMS service deployment at the edge is the bandwidth saving and the relief of the core network from user traffic, a series of comparative evaluation tests were performed, in order to measure the efficiency of the eMBMS technology.

The first test was performed using up to 3 simultaneous UEs receiving the media content either by unicast or by broadcast. The bit rate of the video stream was 1500 Kb/s. The results are shown in Figure 10. In both scenarios the test starts with a single user receiving the video stream and gradually more users receive the service. As depicted in fig. 10, in the case of unicast, the total traffic generated in the S1 backhaul link is increased as more users receive the video stream. On the other hand, in broadcast streaming the traffic remains at the same level no matter how many new users receive the stream.

As it can be observed in Figure 10, during a single user scenario Unicast and eMBMS modes achieve similar performance, in terms of transmitted bit rate, with the eMBMS one, transmitting several kbits/s more. This slight difference is due to the fact that the eMBMS protocol transmits additional packets for synchronization purposes. The SYNC packet traffic generates a constant bitrate of about 50kb/s, independent to the bitrate of the stream, or the number of simultaneous users.

In the graph of Figure 11 the variation of the percentage of the overhead SYNC traffic vs the total stream traffic is presented for a wider range of bitrates from 400-1800 kb/s.

Figure 11 depicts the fact that the overhead of SYNC bytes in an eMBMS service is exponentially decreased as the bitrate of the transmitted stream is increased.

During the experimental process a minor delay in the transmission of the eMBMS stream was noticed and measured. For a set of 3 bitrates (200kb/s, 300kb/s, 500kb/s) for 3 simultaneous users the delay between the video capture and its delivery to the UEs was measured, both for the unicast and the eMBMS modes. The results of the process are presented in Figure 12.

As it can be clearly seen in Figure 12, eMBMS performs slightly worse in terms of content delivery latency. Unicast delivery introduces a latency of about 300ms, whereas eMBMS delivery reaches up to 450ms of delay. Additionally, it must be noted that the total stream delivery is measured, which contains also the encoder's delay, which is approximately 150ms, but it is the same for both the unicast and the eMBMS modes.

From the previous presented set of results, it can be concluded that the trade-off between eMBMS and unicast media delivery for multiple users is between bitrate saving and latency. The latency introduced by eMBMS can be characterized as minor, as it adds to the whole process a mere 150ms delay. When at the same time eMBMS proves to be extremely efficient in terms of resource usage for a large number of users.

VII. Conclusion

This paper presented an NFV enabled Small Cell architecture for 5G systems, which aims to evolve and improve the performance of existing RAN infrastructures. The manuscript not only presented the current issues in the related field, but also discussed viable solutions to these problems.

The paper presented the conformity of the proposed architecture to the evolving 5G New Radio (NR), investigating how the proposed two tier virtualized execution environment can accommodate different deployment options, such as the gNB decomposition between a gNB-CU and one or more gNB-DUs, considered in the 3GPP NG-RAN.

An edge enabler, which has been developed and used in 5G-ESSENCE project, was presented in this paper, which

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

KOURTIS *et al.*: CESC ARCHITECTURE IN 5G NETWORKS FOR BROADCAST/MULTICAST SERVICES 11

efficiently addresses the local break-out issue. Using the SGW-LBO any VNF can be instantiated at the edge (Edge DC in our case), solving the GTP en/decapsulation issue.

The paper investigated the deployment of an MCPTT service on the 5G-ESSENCE platform, where both NFV and MEC capabilities of the platform provided the appropriate environment to guarantee the strict KPIs of mission critical communications. A significant reduction of specific KPIs related to latency was achieved, thanks to the deployment of the MCPTT service at the edge of the network.

Finally, the paper evaluated the performance of a virtualized eMBMS network service deployed at the edge in a media content delivery scenario. The efficiency of eMBMS compared to the unicast method was also investigated. In the future steps the concept of cSON small cells will be investigated in terms of its relevance to RRM and its potential to enhance 5G in terms of performance and efficiency.

REFERENCES

[1] P. S. Khodashenas *et al.*, "Service mapping and orchestration over multi-tenant cloud-enabled RAN," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 4, pp. 904–919, Dec. 2017.

[2] "Cloud-RAN—The benefits of virtualization, centralization and coordination," Stockholm, Sweden, Ericsson, White Paper, 2015.

[3] A. Kostopoulos, I. P. Chochliouros, I. Giannoulakis, A. Kourtis, and E. Kafetzakis, "Small cells-as-a-service in 5G networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–5.

[4] *Mobile Edge Management; Part 2—GS MEC 010-2*. Accessed: Jan. 4, 2019. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/MEC/001_099/01002/01.01.01_60/gs_MEC01002v010101p.pdf

[5] *ETSI Multi-Access Edge Computing (MEC)*. Accessed: Jan. 4, 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/02.01.01_60/gs_MEC002v020101p.pdf

[6] *Mobile-Edge Computing: Introductory Technical White Paper*, ETSI, Sophia Antipolis, France, 2014.

[7] "Network function virtualization (NFV): Use cases, v1.1.1," document GS NFV 001, ETSI, Sophia Antipolis, France, Oct. 2013.

[8] M.-A. Kourtis *et al.*, "T-NOVA: An open-source MANO stack for NFV infrastructures," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 3, pp. 586–602, Sep. 2017.

[9] *Small Cell Virtualization Functional Splits and Use Cases*, Small Cell Forum, London, U.K., 2016.

[10] *Integrated HetNet Architecture Framework*, Small Cell Forum, London, U.K., 2016.

[11] *Rural and Remote Small Cell Network Architectures*, Small Cell Forum, London, U.K., 2015.

[12] *White Papers From the 5G Infrastructure Association*. Accessed: Jan. 4, 2019. [Online]. Available: https://5g-ppp.eu/white-papers/

[13] S. Redana *et al.*, "Views on 5G architecture," 5G-PPP Archit. WG, White Paper, Jul. 2016.

[14] *Multi-Operator and Neutral Host Small Cells: Drivers, Architectures, Planning and Regulation*, Small Cell Forum, London, U.K., 2016.

[15] *Network Functions Virtualization (NFV); Management and Orchestration*, ETSI, Sophia Antipolis, France, 2014.

[16] *ETSI MEC APIS Release*. Accessed: Jan. 4, 2019. [Online]. Available: http://www.etsi.org/news-events/news/1204-2017-07-news-etsi-multi-access-edge-computing-group-releases-a-first-package-of-apis

[17] *General principles for Mobile Edge Service APIs—GS MEC 009*. Accessed: Jan. 4, 2019. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/MEC/001_099/009/01.01.01_60/gs_MEC009v010101p.pdf

[18] *Mobile Edge Platform Application Enablement—GS MEC 011*. Accessed: Jan. 4, 2019. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/MEC/001_099/011/01.01.01_60/gs_MEC011v010101p.pdf

[19] H. Moens and F. De Turck, "VNF-P: A model for efficient placement of virtualized network functions," in *Proc. 10th Int. Conf. Netw. Service Manag. (CNSM) Workshop*, Nov. 2014, pp. 418–423.

[20] A. Gupta, M. F. Habib, P. Chowdhury, M. Tornatore, and B. Mukherjee, "On service chaining using virtual network functions in network-enabled cloud systems," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2015, pp. 1–3. [Online]. Available: http://ieeexplore.ieee.org/document/7413643

[21] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manag. (IM)*, May 2015, pp. 98–106. [Online]. Available: http://ieeexplore.ieee.org/document/7140281

[22] *NG-RAN; Architecture Description; Stage 2 (Release 15), V15.0.0*, 3GPP Standard TS 38.401, Dec. 2017.

[23] "Study on new radio access technology: Radio access architecture and interfaces (release 14), v14.0.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.801, Mar. 2017.

[24] "Study on CU-DU lower layer split for NR (release 15), v15.0.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.816, Dec. 2017.

[25] M.-A. Kourtis, H. Koumaras, G. Xilouris, and F. Liberal, "An NFV-based video quality assessment method over 5G small cell networks," *IEEE Multimedia*, vol. 24, no. 4, pp. 68–78, Oct./Dec. 2017.

[26] P. Salva-Garcia *et al.*, "5G-UHD: Design, prototyping and empirical evaluation of adaptive ultra-high-definition video streaming based on scalable H.265 in virtualised 5G networks," *Comput. Commun.*, vol. 118, pp. 171–184, Mar. 2018.

[27] (Feb. 2018). *ETSI MEC White Paper*. [Online]. Available: www.etsi.org/images/files/ETSIWhitePapers/etsi_wp24_MEC_deployment_in_4G_5G_FINAL.pdf

[28] *MCPTT Over LTE Requirements; Stage 1, v16.1.0*, 3GPP Standard TS 22.179, Apr. 2018.

[29] *Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description, v13.3.0*, 3GPP Standard TS 23.246, Dec. 2015.

[30] M. Fuentes, L. Christodoulou, and B. Mouhouche, "Non-uniform constellations for broadcast and multicast in 5G new radio," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–5.

[31] W. Guo, M. Fuentes, L. Christodoulou, and B. Mouhouche, "Roads to multimedia broadcast multicast services in 5G new radio," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–5.

Authors' photographs and biographies not available at the time of publication.