Enable Advanced QoS-Aware Network Slicing in 5G Networks for Slice-Based Media Use Cases

Qi Wang, Jose Alcaraz-Calero, Ruben Ricart-Sanchez, Maria Barros Weiss, Anastasius Gavras, Navid Nikaein, Xenofon Vasilakos, Bernini Giacomo, Giardina Pietro, Mark Roddy, Michael Healy, Paul Walsh, Thuy Truong, Zdravko Bozakov, Konstantinos Koutsopoulos, Pedro Neves, Cristian Patachia-Sultanoiu, Marius Iordache, Elena Oproiu, Imen Grida Ben Yahia, Ciriaco Angelo, Cosimo Zotti, Giuseppe Celozzi, Donal Morris, Ricardo Figueiredo, Dean Lorenz, Salvatore Spadaro, George Agapiou, Ana Aleixo, and Cipriano Lomba

***Abstract*—Media use cases for emergency services require mission-critical levels of reliability for the delivery of media-rich services, such as video streaming. With the upcoming deployment of the fifth generation (5G) networks, a wide variety of applications and services with heterogeneous performance requirements are expected to be supported, and any migration of mission-critical services to 5G networks presents significant challenges in the quality of service (QoS), for emergency service operators. This paper presents a novel SliceNet framework, based on advanced and customizable network slicing to address some of the highlighted challenges in migrating eHealth telemedicine services to 5G networks. An overview of the framework outlines the technical approaches in beyond the state-of-the-art network slicing. Subsequently, this paper emphasizes the design and prototyping of a media-centric eHealth use case, focusing on a set of innovative enablers toward achieving end-to-end QoS-aware network slicing capabilities, required by this demanding use case. Experimental results empirically validate the prototyped enablers and demonstrate the applicability of the proposed framework in such media-rich use cases.**

***Index Terms*—5G, network slicing, quality of service, eHealth, media use case, verticals.**

Manuscript received October 26, 2018; revised December 10, 2018; accepted February 1, 2019. Date of publication March 13, 2019; date of current version June 5, 2019. This work was supported in part by the European Union H2020 Program through the SliceNet Project under Grant 761913. Parts of this paper have been published in the Proceedings of the IEEE BMSB 2018, Valencia, Spain. *(Corresponding author: Qi Wang.)*

Q. Wang, J. Alcaraz-Calero, and R. Ricart-Sanchez are with the School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, U.K. (e-mail: qi.wang@uws.ac.uk).

M. B. Weiss and A. Gavras are with Eurescom GmbH, 69123 Heidelberg, Germany.

N. Nikaein and X. Vasilakos are with the Communication System Department, EURECOM, 06904 Sophia Antipolis, France.

B. Giacomo and G. Pietro are with Nextworks, 56122 Pisa, Italy.

M. Roddy, M. Healy, and P. Walsh are with the Computer Science Department, Cork Institute of Technology, Cork, T12 P928 Ireland.

T. Truong and Z. Bozakov are with Emerging Tech. and EcoSystem Dev., Dell EMC, Cork, P31 D253 Ireland.

K. Koutsopoulos is with the Systems Engineering Group, Creative Systems Engineering, 11251 Athens, Greece.

P. Neves is with Technology Coordination and Innovation, Altice Labs, 3810-106 Aveiro, Portugal.

C. Patachia-Sultanoiu, M. Iordache, and E. Oproiu are with Development and Innovation/Engineering, Orange Romania, 010665 Bucharest, Romania.

I. G. B. Yahia is with the AI for Network Resiliency and Security, Orange Labs Networks (France), 92320 Paris, France.

C. Angelo, C. Zotti, and G. Celozzi are with Ericsson Telecomunicazioni, 84016 Pagani, Italy.

D. Morris and R. Figueiredo are with RedZinc Services Ltd., Dublin, D08 N9EX Ireland.

D. Lorenz is with Computing as a Service, IBM Research–Haifa, Haifa 3498825, Israel.

S. Spadaro is with Signal Theory and Communications, Universitat Politecnica de Catalunya, 08034 Barcelona, Spain.

G. Agapiou is with the Network Wireless and Core Testing Lab, OTE, 15122 Athens, Greece.

A. Aleixo and C. Lomba are with Protection, Automation and Control, Efacec Energia, 4471-907 Porto, Portugal.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2019.2901402

I. Introduction

LEGACY public safety communications systems, such as Terrestrial Trunked Radio (TETRA), have been designed primarily as private mobile radio systems for the delivery of mission-critical voice, coupled with a select number of narrowband data services, such as text-based messaging. They have been costly to design, deploy and service, and without a fundamental redesign, will not be able to deliver and exploit the media-rich type services currently accessible over public broadband networks.

A selected number of emergency service providers around the globe are deploying public safety communications systems that provide mission-critical mobile broadband services, alongside mobile voice. However, this is an emerging and evolving network environment, with a number of different models being deployed [1]:

- Public safety Mobile Virtual Network Operator (MVNO) (operating or obtaining service from a Secure-MVNO) - In 2015 the U.K. Home Office awarded a £1 billion Emergency Services Network (ESN) contract to network operator EE and Motorola Solutions to migrate from the dedicated TETRA network to a public safety S-MVNO over a public broadband Fourth Generation (4G) network. The original migration date was due to start in 2017 [2] but continued rollout delays have caused the Home Office to rethink their strategy, by taking a phased approach, with initial services not likely to come on stream before end of 2019.

- Dedicated (building a dedicated network) - Qatar has deployed a dedicated nationwide 4G network, specifically as a public safety communication system.
- Mobile Network Operator (MNO) (contracting services through an existing MNO) - The New York Police Department (NYPD) has equipped their police force with mobile devices connected to the public network.
- Hybrid solution (combining a public safety MVNO with a dedicated network) - The U.S. FirstNet public safety network operates as a hybrid over dedicated and public networks.

Public safety agencies have specific requirements in terms of Quality of Service (QoS). When mission-critical services are carried over public broadband networks, exact levels of security, resilience and reliability must be guaranteed. This is the challenge, and the Horizon 2020 5G Public-Private Partnership (5G PPP) SliceNet project [3] is investigating this challenge by designing the seamless deployment of media-rich public safety mobility services over future 5G broadband networks, by employing an advanced network slicing framework to ensure mission-critical QoS.

There is rapidly growing recognition of the significance of network slicing to enable differentiation of various services with diverse QoS requirements such as mission-critical services and best-effort services. Major European and worldwide 5G industries have declared the importance of achieving network slicing as a fundamental architectural requirement and critical enabler of 5G networks [4]. Moreover, the last couple of years have witnessed an increasing number of headlines, highlighting early trials of network slicing by leading operators or vendors.

Despite these encouraging efforts, there are a number of gaps that need to be filled to fully realise the envisioned benefits of 5G slices. Firstly, end-to-end (E2E) network slicing has been specified by the International Telecommunication Union (ITU) [5] as a high-priority technical gap to tackle for E2E application quality. In particular, plug and play (P&P) customisation of network slicing for verticals has yet to be achieved. Secondly, recent vertical markets' analyses and related studies [6] have shown the importance of forming a close partnership between 5G industry and vertical business sectors in achieving the fully connected society vision in 5G. However, previous 5G projects have mainly taken a technology-driven approach and have focused on the enabling 5G technologies, whilst only the latest 5G projects have started to emphasise a business-driven approach to promote 5G. It is thus essential to offer a cost-effective migration pathway and a one-stop interface for verticals (e.g., regarding the concerned emergency services) to adopt 5G slice-based/enabled services. Finally, the QoS assurance in 5G slices has yet to be adequately addressed, especially for media-centric use cases such as eHealth. Upon achieving slicing, a major challenge is to maintain or optimise the quality for the vertical businesses. QoS should not be achieved via over-provisioning of resources, which is expensive and not scalable.

Motivated by the above context, SliceNet aims to drive 5G network slicing to the next level, by pushing the boundaries significantly in fulfilling the challenging requirements from the management and control planes of network slicing across multiple administrative domains, facilitating early and smooth adoption of 5G slices for verticals to achieve their demanding use cases, and managing the QoS for sliced services. To realise this highly ambitious vision, a novel and powerfully integrated network management, control and orchestration framework is proposed for adoption of 5G network slicing and slice-based services. The SliceNet framework is designed to be applicable to a wide range of vertical businesses, and in this paper we focus on its applicability to media-centric use cases, where demanding QoS requirements, especially low latency, are expected in the context of the SliceNet eHealth use case. The novel contributions of this paper can be highlighted as follows.

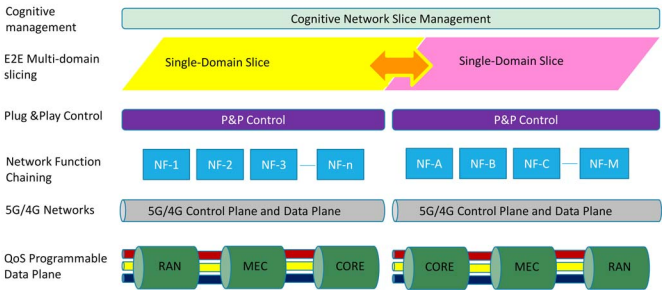


Fig. 1. Network slicing approaches in SliceNet.

- Design of a QoS-aware network slicing framework for media and other vertical industries with diverse QoS requirements.
- Design and prototyping of a set of innovative enablers to allow efficient (through a One-Stop Application Programming Interface (API)), customisable (through flexible and on-demand P&P Control), QoS-assured network slicing (through exploring advanced resource allocation and data plane programmability) for media use cases in an open source based 5G network slicing infrastructure (OpenAirInterface [7] and Mosaic-5G [8]).
- Experimental results of key enablers to demonstrate the performance to meet the challenging requirements in a 5G eHealth case study.

The remainder of the paper is organised as follows. Section II provides an overview of the SliceNet technical approaches and a media-centric eHealth use case. Section III elaborates the eHealth use case by providing detailed design of a set of key enablers for this use case, and Section IV describes the prototyping of these enablers with implementation details, and presents experimental results obtained from the prototyping tests. Finally, Section V concludes the paper and outlines future work.

II. Overview of the SliceNet Approach

SliceNet aims to design, prototype and demonstrate an innovative, verticals-oriented network slicing framework in Software-Defined Networking (SDN) and Network Function Virtualisation (NFV) enabled 5G networks. The novel network slicing approaches in SliceNet are highlighted in Fig. 1,

emphasising the different levels of essential and advanced network slicing.

Firstly, the network function chaining enables best-effort slicing, which is the state-of-the-art essential network slicing approach, follows a network function forwarding graph to achieve specific network services for a use case slice. The network functions can be Virtual Network Functions (VNFs) or Physical Network Functions (PNFs). Secondly, SliceNet advances the state of the art through QoS-aware slicing, which attempts to realise network slicing with guaranteed network-layer QoS (e.g., in terms of bandwidth, delay/latency etc.), to be achieved by the QoS-programmable data plane across the various network segments. To this end, SliceNet introduces QoS-enabling mechanisms such as hardware-based or software-based traffic engineering functions in the network. This is in line with the vision from 3GPP to "Provide slice-as-a-service with guaranteed QoS" [9]. Thirdly, through a novel P&P Control layer, SliceNet can achieve customisable network slicing, which allows a vertical user to request on-demand control functions to be enforced regarding a slice for the vertical, depending on the levels of exposure of runtime network control capabilities from the operator. Conceptually, it can be a push from the vertical to insert a custom P&P Control function to the slice or a pull, e.g., in terms of monitoring functions exposed by the operator. Fourthly, SliceNet targets E2E network slicing. In the first phase reported in this paper, E2E refers to the path from the Radio Access Network (RAN), via the Mobile/Multi-access Edge Computing (MEC) platform to the Core Network (CN) within a network service provider's domain. In the later phase of the project, E2E will extend to multi-domain scenarios. Finally, SliceNet will introduce cognitive network slice management to achieve Quality of Experience (QoE) aware slicing, which further advances QoS-aware slicing yet is out of the scope of this paper.

One of the primary use cases in the project to validate and demonstrate the benefits of the SliceNet technical approaches is the eHealth Smart/Connected Ambulance Use Case. The Connected Ambulance will act as a connection hub for the emergency medical equipment and wearables, enabling real-time streaming to the awaiting emergency department team at the destination hospital. This use case will advance the emergency ambulance services by developing new collaborative models with the healthcare stakeholders to help create improved experiences and outcomes for patients in their care. Detecting the demeanour of patients in real-time can provide a basis for enhanced care and can alert medics to situations where intervention is required. A patient's demeanour can indicate health conditions ranging from discomfort to severe events such as stroke. The eHealth use case leverages MEC for emergency telemedicine over 5G networks. Dedicated 'slices' of the network are guaranteed to ensure the QoS necessary for the delivery of mission-critical services, with the help of data plane programmability. These services can offer powerful applications that can provide life-saving diagnostics by utilising high resource availability on the MEC. This approach reduces the need for high computing resources in every ambulance.

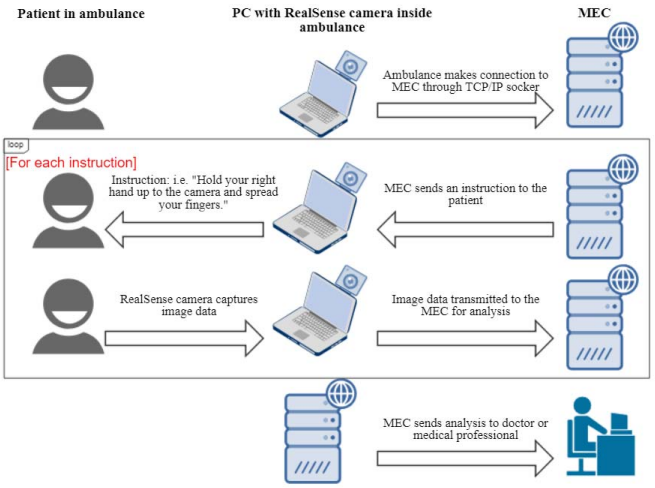


Fig. 2. High-level architecture for TeleStroke Assessment service.

III. Network Slicing Based Multimedia Communications: eHealth Use Case Study

A. eHealth Use Case Scenario

Stroke is the number one cause worldwide of acquired disability, number two cause of dementia, and number three cause of death. In the U.K. alone, the disease costs the National Health Service £9 billion a year [10]. To date, significant progress has been made on in-hospital stroke management, but a reliable pre-hospital in-ambulance solution has not yet been established. Solving this problem could offer life-saving time gains and speed up treatment initiation by early activation of the in-hospital stroke response, thereby curtailing the risk of misdiagnosis and death.

Generalised in-ambulance telemedicine is a recently developed and promising approach and SliceNet exploits MEC-enabled network slicing, together with Machine Learning (ML) to facilitate widespread use of an in-ambulance telestroke diagnostic service that could radically improve future patient treatment pathways.

The state of the art in pre-hospital care consists of basic paramedic procedures before the patient arrives at the hospital. After arriving, the patient goes through standard procedures and tests before reaching a final diagnostics and a specific therapy can begin. By utilising time spent in transit within an ambulance, pre-hospital screening can provide an initial diagnostic for medical professionals who can then investigate further to confirm the diagnosis and begin specific stroke therapy. This offers a significant time gain since the patient arriving at hospital has already completed an initial triage screening. In addition, some emergency journeys in rural areas from incident location to hospital can be long, and early diagnostics in-ambulance have the potential to increase emergency procedures pre-hospital, thereby improving patient outcome.

The main QoS requirements in the eHealth use case are 10 Mbps or higher throughput (per video slice) and up to 30 ms E2E latency. In addition, it requires high reliability, wide area coverage and mobility support.

B. System Design

SliceNet has designed a MEC-based TeleStroke Assessment service (Fig. 2), which can be accessed on demand by paramedics in the ambulance, to assess symptoms of patients suspected of suffering from a stroke. The assessment is derived from an existing Unassisted TeleStroke Severity Scale (UTSS), which would normally involve a series of clinician delivered steps to the patient in the hospital. The application gives the consultant located at the hospital a first indicator of the type of stroke the patient might have, and thus enables them to make the necessary clinical arrangements ahead of patient arrival at hospital.

The application is connected to a cloud program running on the MEC, which is responsible for sending instructions to the patient and to perform analysis on video data, sent from the ambulance application. Instructions used to conduct the examination are defined in the UTSS protocol, to detect possible signs of stroke. The MEC application sends the first instruction to the patient in the ambulance, which in turn starts transmitting voice and image data of the patient to the MEC. As the MEC receives the data it performs an analysis using ML to detect for possible signs of stroke and sends the results of the analysis to the hospital/doctor.

C. Key Technical Enablers and Innovations

1) Network Slicing: Network slicing is one of the key enablers to flexibly deliver networks on an as-a-service basis. It enables the composition and deployment of multiple logical networks over a shared physical infrastructure, and their delivery as a service or slice. A slice can either be completely isolated from other slices, down to the different sets of spectrum and cell site (as in most of the current 3G and 4G deployments), or be shared across all types of resources including radio spectrum and network functions (e.g., all network layers of the protocol stack). Another alternative is for a slice to be customised for a subset of User Plane (UP) and Control Plane (CP) processes with an access to a portion of radio resources in a virtualised and/or physical form. 3GPP provides several studies related to the E2E network slice management and orchestration in TR 28.801, service-oriented 5G core network in 3GPP TS 29.500 and TS 23.50, and RAN slicing aspects in TR 38.801. However, RAN slicing beyond well-known 3GPP sharing models, as well as the interplay with a service-oriented core network to enable E2E network slicing, remains challenging, particularly in terms of different levels of isolation and sharing of network resources and state. This calls for a unified and flexible execution environment to run multiple virtualised RAN and CN instances with the required levels of customisation over heterogeneous deployments. To this end, we present below the proposed runtime slicing system [11].

The E2E network slicing architecture is shown in Fig. 3 with the RAN and CN runtime being the core components by which each slice interacts with underlying RAN and/or CN function to access resources and states as well as to control the underlying network behaviors.

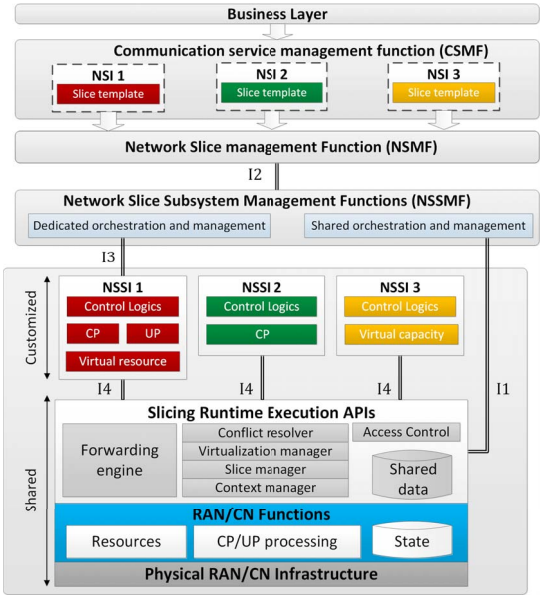


Fig. 3. RAN and CN runtime slicing system.

There are five services provided within the slicing runtime execution environment: (a) slice manager, (b) virtualisation manager, (c) context manager, (d) conflict resolver, and (e) access control, all operating over a shared slice data. Slice data includes both slice context (e.g., basic information to instantiate a slice service such as its identity, user context and their slice association) and module context (e.g., CP and UP state information, RAN and CN function primitives and APIs) that are used to customise and manage a slice in terms of the required resources, states, processing, and users. Slice data can be transferred or shared among different runtime instances dynamically due to the user and network dynamics, e.g., user handover and/or service template change when updating the functional splits.

Based on the slice service template and slice context, the slice manager determines the CP/UP processing chain for each slice and each traffic flow, and programs the forwarding plane allowing directing the input and output streams across multiplexed processing operated by the underlying RAN and CN functions and customised processing performed by each slice. An E2E service life-cycle is operated by the slice manager in support of service continuity when the slice service template is updated. Based on the service definition and slice context, the slice manager determines the CP/UP processing chain for each slice and each traffic flow, and programs the forwarding engine through a set of rules to direct the input and output streams across the multiplexed processing operated by underlying network infrastructure function and the customised processing performed by the slice. Moreover, the slice manager is responsible to apply Service Level Agreement (SLA) policies, detect and resolve conflicts, and control access to resources and state. The virtualisation manager provides the required level of isolation and sharing among slices. Specifically, it partitions resources and states, abstracts resources and states to/from the virtual ones, and reveals virtual resources and states to a slice, which are decoupled from

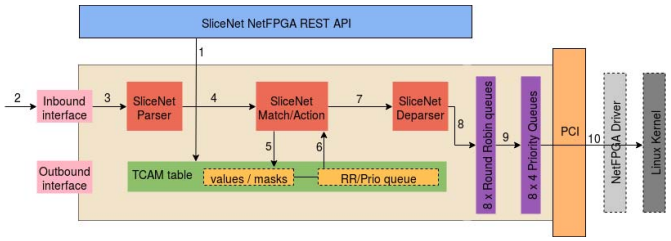


Fig. 4. SliceNet data plane traffic classification and QoS control architecture.

the physical ones. The context manager performs CRUD operations (i.e., create, read, update, and delete) on both slice and module context. To create a slice context, it performs slice admission control based on the service template that defines the required processing, resources, and states. Upon admission control, module context is used to register (a) slice-specific life-cycle primitives to the slice manager, and (b) requested resources and/or performance to the virtualisation manager.

To facilitate slice orchestration and management, four interfaces (I1 to I4) are provided between the slice runtime and the communication service management function (CSMF), network slice management function (NSMF) and network slice subnet management function (NSSMF) defined by 3GPP in TR 28.801.

2) QoS Control Based on Data Plane Programmability: Fig. 4 presents the design of a hardware-accelerated programmable data plane, suitable for the transmission of multimedia applications with controllable QoS. The main architecture for prototyping is based on NetFPGA (10 Gbps) cards and the P4 programming language, following the SimpleSumeSwitch architecture [12]. The proposed architecture in this paper allows not only traffic parsing and classification based on the solution presented in [13], but also applying QoS control to the multimedia especially video transmission over 5G networks. This QoS control is applied through a Representational State Transfer (REST) API, which allows selecting the priorities of the 5G network traffic processed by the NetFPGA. A QoS-aware P4-based firmware has been developed for this data plane, implementing a double level of queuing. The first level comprises eight round robin queues, which will ensure a proper balance between the traffic processed. The second level further deploys four priority queues per each of the previous round robin queue. This second level of queues will allow achieving a significant and constant low latency for mission-critical network traffic such as the video streaming based emergency communication in the concerned eHealth use case.

Fig. 4 shows the following sequence that describes the behaviour of a programmed NetFPGA card when an inbound 5G network packet arrives at the NetFPGA with the proposed QoS-aware firmware loaded:

1) The traffic classification rules are inserted in the NetFPGA's Ternary Content Addressable Memories (TCAM) table through the REST API.
2) An inbound packet arrives at the NetFPGA through a physical interface.
3) The packet is sent to the Parser for classification.
4) Once the packet has been classified based on its headers, it is sent to the Match/Action component.
5) In the Match/Action, the TCAM table is checked to see if any special priority should be applied to the inbound packet.
6) If there is any rule that matches the packet, the second part of the rule will be received by the Match/Action (step 6). The second part of the rule contains the round robin and priority queue where the packet will be sent.
7) The Match/Action applies to the packet the slicing configuration received by the second part of the rule (step 6) and then the packet is sent to the Deparser.
8) The Deparser builds the packet to be sent to one of the eight Round Robin queues indicated in the rule.
9) The packet is sent to one of the four priority queues. If there is no rule that matches the packet, it will be sent to the lowest priority queue.
10) Packet leaves the priority queues and is sent to the Linux Kernel through the Peripheral Component Interconnect (PCI) and the NetFPGA driver.

It is noted that there are major differences of network slicing from DiffServ. In the context of the proposed QoS-aware network slicing based on the data plane programmability, effectively a packet scheduler is achieved to ensure that the different traffic flows will not affect each other in terms of performance, referred to as performance isolation, which is essential to warrant the performance of different slices. Within each of the slices, each can have DiffServ though. Moreover, DiffServ only works in traditional pure IP networks but not in overlay networks with encapsulation protocols such as those in 5G networks. In addition, DiffServ only defines well-known semantics on how to deal with the traffic but not the programmable way to define such semantics.

3) Low Latency MEC Platform: MEC, being a cloud-based service environment at the edge of the network, offers real-time, high-bandwidth and low-latency access to radio network information. This allows applications at the network's edge to monitor and control the underlying networks in the close proximity of their users. The proposed Low Latency MEC (LL-MEC) platform [14], [15], in particular, is the first open-source 3GPP-compliant implementation covering multiple APIs aligned to ETSI MEC specifications. LL-MEC uses the extended OpenFlow [16] and FlexRAN [17] protocols, addressing three types of latency: (i) E2E user transport latency (i.e., "User latency"); (ii) Control latency, capturing the underlying network latency for MEC to perform an action on behalf of an edge application; and (iii) Application latency for performing edge application actions.

Fig. 5 shows a high-level view of the platform architecture, operating on a software-defined mobile network that consists of multiple Long-Term Evolution (LTE) eNBs and physical or software OpenFlow-enabled switches. At the top, the application manager founds the upper-most layers, providing the API (Mp1) for applications. The middle layer includes the Radio Network Information Service (RNIS) and the Edge Packet Service (EPS) that manage RAN and CN services, respectively, based on the CP and UP APIs in the abstraction layer,

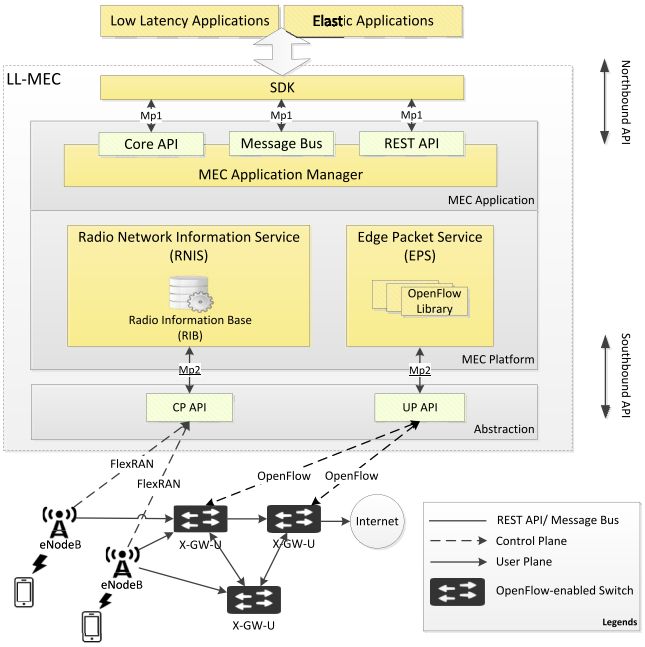


Fig. 5. High-level schematic of LL-MEC.

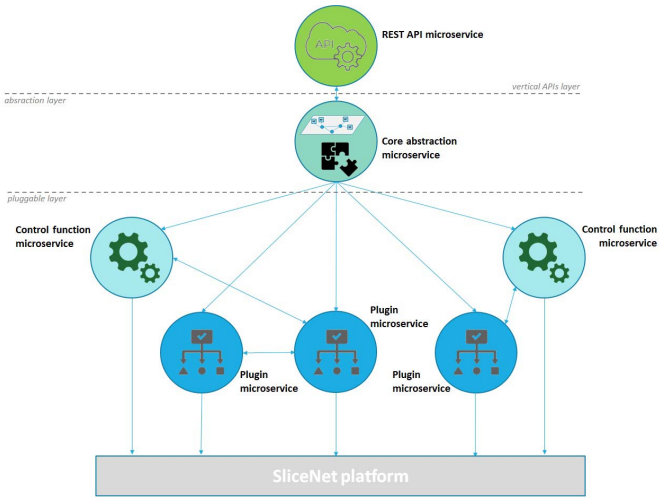


Fig. 6. P&P microservices architecture.

respectively. Below the abstraction layer, eNBs and OpenFlow-enabled switches comprise the UP functions, with FlexRAN and OpenFlow comprising the CP functions, abstracting all information and exposing it via the abstraction API (Mp2).

4) One-Stop API: The SliceNet One-Stop API facilitates all the potential user roles to be provided with different views and different control and management options tailored to their usability and workflow needs. Consequently, the identity of the user utilising the One-Stop API user interface is subject to role analysis for the resolution of the access rights across the platform functional endpoints and information sets (including inventory and monitoring data) as these are stemming from the SLAs and the related foreseen privileges. In practice, the instance of a slice template, as it can be expressed by the relationships among instances of service components, network functions and infrastructure resources, is annotated with access rights options that are taken into account when a user's role is applied. This process leads to the composition of the operation space through which the user interacts either with the slice and service instances as a vertical user, or with the resource instances ranging from physical to logical and service resources in the role of a provider (network or digital service provider). While the vertical view aims at covering the application-specific requirements that a slice is intended to support, the provider role views are utilised in the process of designing network or digital service offerings, which will be selected and utilised by the consumer role (network or digital service consumer respectively) on top of it. At the interfacing border between the role of each provider with the role of the related consumer, offerings are abstracted to be conceived by the consumer role's design and/or operation processes, and conversely consumer requests are expanded and handled in more complex workflows for the delivery of the higher level requirements. This variation in slice and service view and control exposure has been recently exercised by benefiting from the following P&P Control framework as an enabling control framework towards provisioning customised views.

5) P&P Control: One significant gap in the current network slicing systems is that verticals and slice consumers' role and involvement in the runtime control and management of their slices is very limited. The SliceNet P&P Control aims at going beyond this current practice by enabling a customisation process of the slice exposed control. Dedicated vertical-tailored slice control instances are exposed and offered to verticals through the P&P Control framework, as isolated control environments where customised control logic and functions are plugged to build a truly tailored slice view. The idea is that each vertical may have different requirements on how they would like to control, manage or monitor their slices, in terms of exposed (topological) view and runtime operations they could enforce. The P&P Control fulfils these requirements by providing a common framework, which through a tight integration with the One-Stop API layer, is able to expose customised slice control views and operations.

The P&P Control covers two fundamental aspects: firstly, how the slice is exposed, i.e., how it is presented to verticals and slice consumers in terms of components, network and service functions, and topology; secondly, how the slice and its components can be controlled and managed at runtime, i.e., which tailored and customised runtime operations can be applied by vertical and slice consumers. One of the main and key features enabling this two-fold flexible control exposure is the P&P design approach based on microservices. Each P&P instance is a collection of self-contained microservices, as depicted in Fig. 6. This enables a truly dynamic and flexible approach, as P&P components can be activated or updated at runtime without disrupting the overall P&P Control instance integrity and continuity, in particular when selecting proper technologies for containerisation and orchestration of microservices (e.g., Docker [18] and Kubernetes [19]).

The control function and plugin microservices are self-contained functions implementing vertical-tailored control

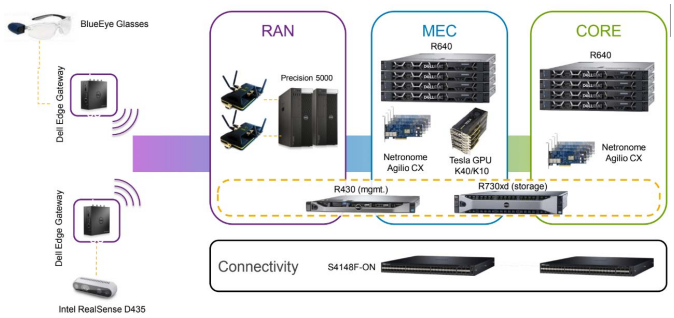


Fig. 7. 5G eHealth prototyping - hardware stack.

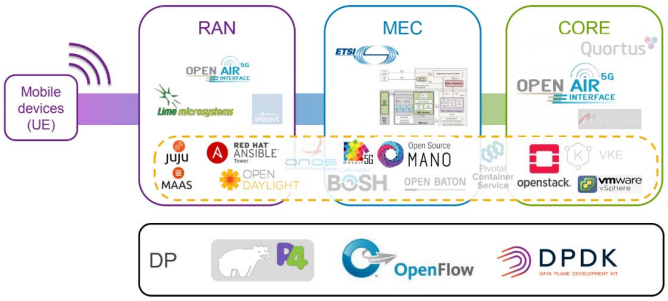


Fig. 8. 5G eHealth prototyping - software stack.

logic, thereby leveraging and coordinating APIs and data collected from the SliceNet platform, while exposing to the core abstraction microservice (and thus to verticals) new APIs and logic not natively offered by SliceNet. Therefore, each P&P Control function allows further specialising the dedicated slice control environment offered to verticals in order to meet the required level of control exposure agreed. In the case of eHealth, a key control capability and logic exposed towards the vertical is related to the deployment of the Telestroke VNF at proper and vertical-selected MEC locations. The core abstraction microservice is the core engine of each P&P Control instance, and it represents the main enabler of the slice control exposure customisation, in terms of both vertical-tailored slice view (i.e., network and service functions components) and offered runtime control operations. In particular, it translates and abstracts the specific APIs, logic and data offered by the plugin and control function microservices into vertical-tailored operations. To this end, it manages in a fully dynamic and flexible way a technology-agnostic slice exposure model that allows representing a given slice instance as a collection of slice elements organised in a topology, augmented with a set of control capabilities, attributes, operations and primitives exposed on top of it. The instantiation of this model is specialised by the P&P Control for each slice instance according to the vertical requirements and business logic, assigning to each slice element in the vertical-tailored slice context. In the case of the eHealth use case, this slice topology view allows exposing details related to main slice components, such as RAN, MEC and CN features and locations, on top of which the vertical can apply its runtime logic, e.g., for deploying a custom Telestroke VNF.

IV. eHealth Use Case Prototyping and Experimental Results

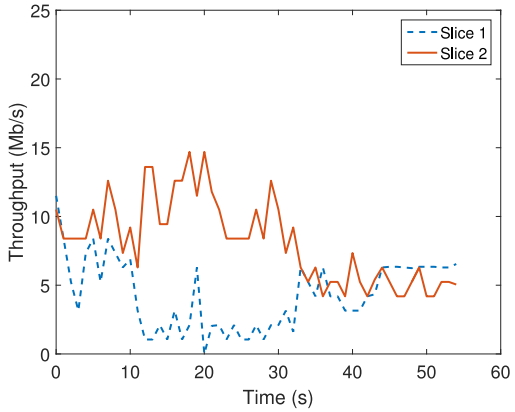
A. Infrastructure and MEC App

Fig. 7 and Fig. 8 show the hardware and software stack of the 5G Prototyping Lab at Dell EMC facilities in Cork, Ireland. The Telestroke VNF has two parts. Firstly, it has a gateway application which is running on the Dell Edge Gateway. The gateway application is distributed to ambulances and requires an Intel Realsense camera to capture the video of a patient performing the UTSS. The UTSS features a 14-step process which is recorded by the camera and those image frames are sent over a Websocket, using 5G technology, to a server. Secondly, it has a server running at the MEC segment as a MEC App, which will accept connection from the gateway and start receiving the images. Each image frame is analysed by the server's ML-based telestroke assessment algorithms and outputs meaningful diagnostic information to medical professionals, for example the stroke condition of the patient.

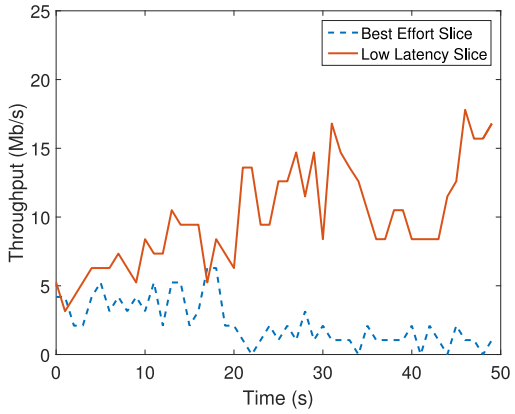
B. RAN-Core Slicing and MEC Platform

To demonstrate the benefits of E2E slicing, we apply two different RAN resource allocation strategies and implement a low-latency MEC application interfacing with the LL-MEC platform via the northbound Software Development Kit (SDK) (Fig. 5). We change the enforced policy on-the-fly and measure the downlink throughput in two scenarios: using (a) an *uncoordinated* and (b) a *coordinated* E2E resource programmability scheme in terms of radio resources for the RAN and switching bandwidth for the CN. The rationale behind the choice of these two extreme schemes is to demonstrate the advantages of coordinated E2E resource programmability in comparison with uncoordinated programmability, with respect to resources allocation efficiency and meeting the exact performance requirements of slices. The corresponding performance results appear in Fig. 9. For the case of uncoordinated programmability, a policy is applied at time $t = 10s$ according to which slice 1 must be allocated with 1 Mbps and slice 2 15 Mbps. A second policy is applied at $t = 20s$, but only to the RAN part, so as to lower the rate down to 50% of the radio resources, i.e., to 8 Mbps for each slice. Finally, we enforce a third policy only to CN at $t = 33s$ to increase the switching bandwidth to 6 Mbps, whereas for coordinated programmability only one policy is enforced at $t = 18s$ to both the RAN and the CN, hence creating a best-effort slice with 1 Mbps and a low latency slice with 15 Mbps, meeting the QoS requirement of 10 Mbps or higher throughput for this eHealth use case.

The benefits of MEC and its unified SDK for enabling coordinated programmability and network slicing are evident. For the case of (a) uncoordinated slicing the bandwidth is used inefficiently due to the asynchronous and uncoordinated allocation of resources between RAN and CN. For (b) coordinated slicing, however, the anticipated performance gap between the "Low-Latency Slice" and the "Best-Effort Slice" is clear, as the resources are allocated to each slice according to their specific requirements.



(a) Uncoordinated slicing



(b) Coordinated slicing

Fig. 9. Mobile network slicing.

C. QoS-Aware Data Plane for Slicing

To achieve QoS-aware data plane in the 5G non-RAN segments (e.g., MEC, MEC to CN, and CN), we have prototyped a slicing-friendly infrastructure especially a programmable data plane for 5G MEC infrastructure. The prototyping utilised the P4 NetFPGA reference implementation recently released by NetFPGA employing the Xilinx SDNet P4 compiler. The SDNet Compiler v2017.1.1 [20] was set up on Ubuntu (64-bit) with Vivado Design Suite installed and licensed, and also Xilinx SDK 2016.4 [21]. The main NetFPGA features used for the prototyping are as follows: a Xilinx Virtex-7 FPGA, one PCIe Gen3 x8, a Xilinx CPLD XC2C512, three x36 72Mbits QDR II SRAM, two 4GB DDR3, one Micro USB cable for programming/UART, and four SFP+ 10Gbps interface.

Experimental tests have been conducted to empirically validate the design and prototyping. In the tests, the following video-rich scenario has been chosen. We programmed the NetFPGA card to be in the NIC Mode with support for 5G network slicing traffic. Four different types of video slices were considered including ultra-low-latency eHealth video communication, low-latency video conferencing, medium-latency video surveillance and best-effort video entertainment. eHealth traffic is the first priority traffic and was sent through Slice 3, which provides the lowest latency. On the other hand, best-effort video traffic was sent through slice 0, which has no guaranteed QoS for the network traffic. We have emulated

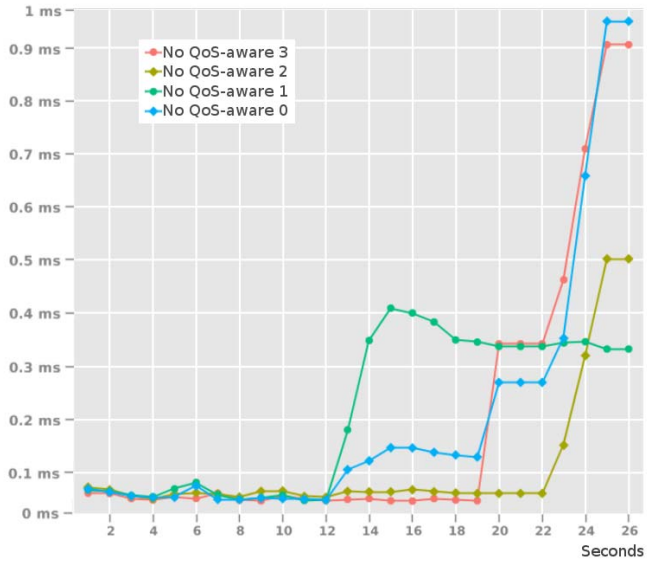


Fig. 10. 256 users transmitting H.265 video without QoS-aware network slicing.

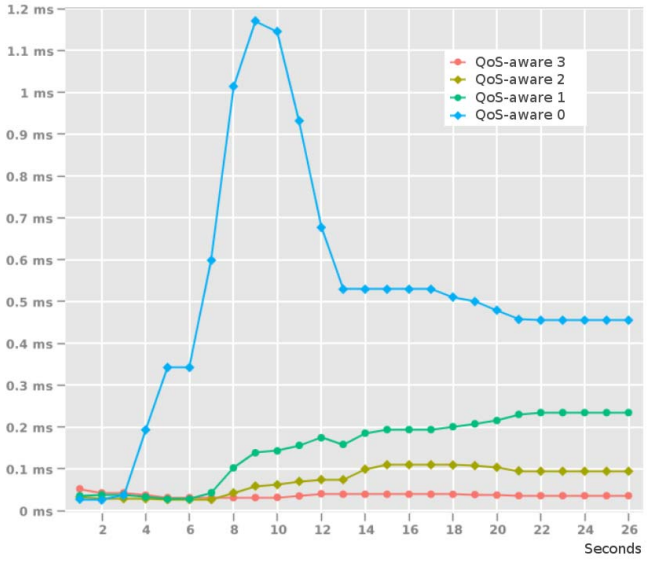


Fig. 11. 256 users transmitting H.265 video with QoS-aware network slicing.

64 users per slice, transmitting H.265 encoded video simultaneously. Therefore, a total of 256 users were emulated for the four slices. The average of the bandwidth use for the H.265 video transmission was 1.07 Gbps, the transmission of the video for the experiments took 26 seconds, and thus the total data transmitted in the 26 seconds was 27.8 Gb.

Fig. 10 shows the average per-hop delay performance for H.265 video transmission of 256 users belonging to four slices whilst there was no QoS-aware network slicing applied to the network traffic. As can be observed, the delay behaviours of the four best effort slices are random, and thus there is no differentiation between the four groups of users who are transmitting video traffic at the same time. This scenario does not guarantee a low latency for any network transmission, thereby posing a real hurdle for mission-critical traffic in an eHealth use case.

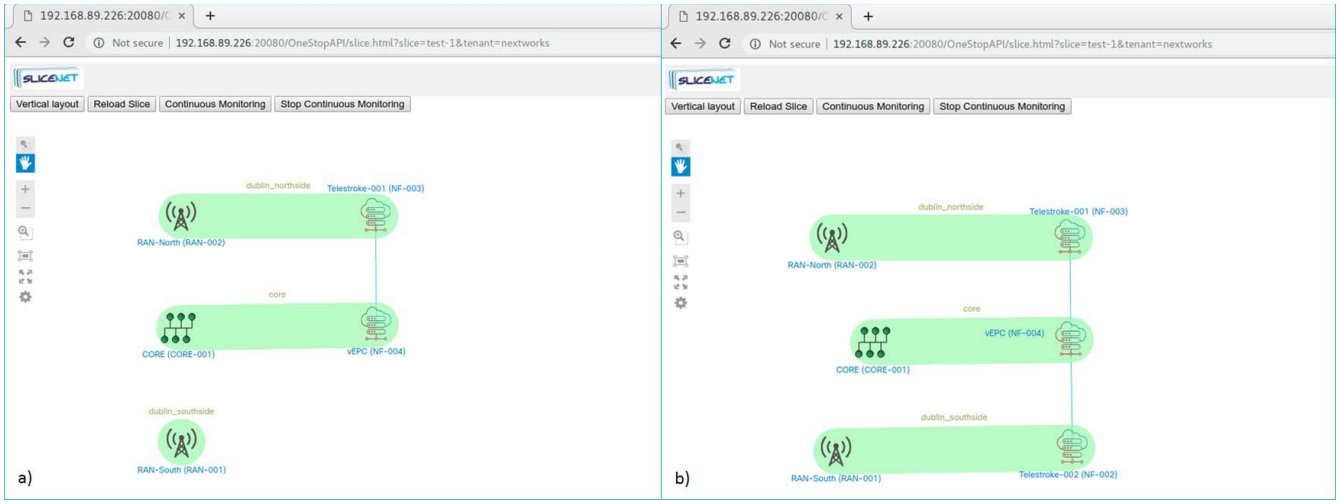


Fig. 12. Customised eHealth slice topology view as exposed by the P&P Control through the One-Stop API GUI.

In contrast, Fig. 11 shows how the proposed QoS control in the data plane allows the infrastructure to fulfil the QoS differentiation of different network traffic. As shown in Fig. 11, the delays for the eHealth traffic (QoS-aware 3, in red) remained constantly the lowest during the video transmissions of the 64 users belonging to that slice with top priority, with an average delay of less than 0.05 ms. In comparison, it is clear that the 64 users who were sending video traffic with QoS-aware 0 (in light blue) experienced significantly higher latency throughout the course, because that group of users did not have any slicing priority applied.

It is noted that this delay is a per-hop measurement of the latency for proof-of-concept in the prototyping. In an E2E network involving multiple domains, this per-hop latency would add up to noticeable values to be perceived by the video application users. Meanwhile, it is expected that the up to 30ms E2E delay performance requirements in this use case will be met in a multi-domain setting of an integrated testbed.

D. P&P Control and One-Stop API

The P&P Control prototype realises a dedicated per-slice control instance, tuned to the vertical's requirements. Each P&P Control instance consists of a specific set of Docker containers, properly selected in order to provide all the features needed to build a control panel that meets the vertical's requirements. The containers are orchestrated by Kubernetes, which is in charge of managing the life-cycle of each P&P Control instance, by exposing it as Kubernetes service. Such a service exposes the P&P Control interface for management and service to verticals, and it is fully compatible with the One-Stop API Graphical User Interface (GUI).

Through the GUI, as shown in Fig. 12 a), a vertical can exploit a graphical view of the resources belonging to its own slice and interact with them, according to the functionalities provided by a proper P&P Control instance. In this specific case, the P&P Control and the One-Stop API have been integrated and validated in the context of the eHealth use case. Fig. 12 shows a set of VNFs (vEPC, Telestroke App etc.), connected in the form of a graph. In response to the vertical's requirements, the P&P Control interface can provide the feature of registering a new resource to the slice, as shown in Fig. 12 b), which represents the status after an additional Telestroke App is added. Furthermore, the P&P Control allows the vertical to activate continuous performance monitoring feature at different levels (slice-level and user equipment level, e.g., ambulance throughput) and at different granularity degrees by offering the possibility of creating data stream channels, based on WebSockets sessions between the P&P Control instance and the One-Stop API GUI, as shown in Fig. 13.

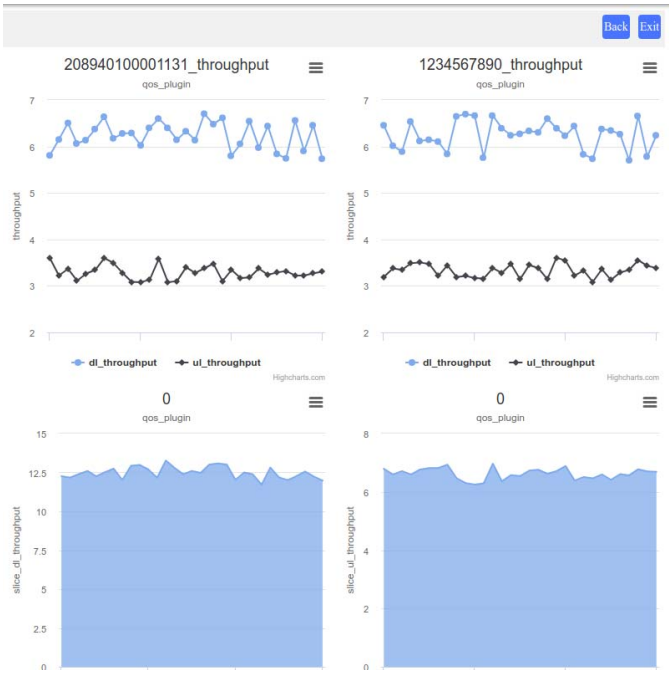


Fig. 13. P&P Control exposed slice and user equipment throughput continuous monitoring information (shown on the One-Stop API GUI).

V. Conclusion and Future Work

The emergence of the promising 5G networking has motivated a major paradigm shift for eHealth services from expensive dedicated private healthcare networks to affordable 5G public cellular networks. Meanwhile, eHealth services are typically media-rich and mission-critical, demanding high degree of QoS guarantee, comparable to that in the private healthcare counterpart.

This paper has presented the SliceNet approach to meeting such challenging requirements through an advanced network slicing framework over a programmable 5G infrastructure. A number of innovations and technical contributions from SliceNet have been described, focusing on the design and prototyping of a set of key technical enablers for a media-centric eHealth use case, which features ML-based early diagnosis of patients in an emergency scenario and QoS assurance for real-time uplink video streaming from an ambulance to the hospital specialist. These key SliceNet technical enablers for such media use cases include a One-Stop API to facilitate a vertical view of their slice, RAN and Core slicing to meet the use case's specific resource requirements, a low-latency MEC platform to allow speedy media processing, hardware-accelerated QoS-controllable data plane, and a P&P Control framework to allow runtime customisation of the slice. Prototyping and experimental results have demonstrated the applicability of these SliceNet enablers in such media-centric use cases.

Ongoing work focuses on the integration of the presented technical enablers in a complete SliceNet framework, and conducts further empirical tests to gain more insightful understanding at the system level. Further work is underway on cognition algorithms to improve the QoE of the eHealth use case.

Acknowledgment

The authors would like to thank all SliceNet partners for their support in this work.

References

[1] NOKIA. (2016). *Four Business Models for Mobile Broadband Public Safety Communications*. [Online]. Available: https://onestore.nokia.com/asset/182917

[2] S. Mccaskill. (2018). *Government 'Unlikely' to Scrap Delayed 4G ESN*. [Online]. Available: https://www.techradar.com/news/government-unlikely-to-scrap-delayed-4g-esn

[3] Q. Wang *et al.*, "SliceNet: End-to-end cognitive network slicing and slice management framework in virtualised multi-domain, multi-tenant 5G networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–5.

[4] O. Y. Bellégo. (Sep. 2016). *Orange 5G Vision*. [Online]. Available: https://www.orange.com/fr/content/download/38314/1162326/version/2/file/Orange%205G%20vision%20(en%20a%20nglais,%20septembre%202016).pdf

[5] International Telecommunication Focus Group IMT-2020. (Dec. 2015). *Report on Standards Gap Analysis*. [Online]. Available: https://www.itu.int/es/ITU-T/focusgroups/imt-2020/Pages/default.aspx

[6] 5GPPP. (Feb. 2016). *5G Empowering Vertical Industries*. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

[7] N. Nikaein *et al.*, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 33–38, Oct. 2014.

[8] N. Nikaein, C.-Y. Chang, and K. Alexandris, "Mosaic5G: Agile and flexible service platforms for 5G research," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 48, no. 3, pp. 29–34, Jul. 2018.

[9] 3GPP. (2018). *Study on Management and Orchestration of Network Slicing for Next Generation Network (Release 15)*. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/28_series/28.801/28801-f10.zip

[10] NCBI. (2009). *Cost of Stroke in the United Kingdom*. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19141506

[11] C.-Y. Chang and N. Nikaein, "RAN runtime slicing system for flexible and dynamic service execution environment," *IEEE Access*, vol. 6, pp. 34018–34042, 2018.

[12] N. Zilberman *et al.*, "NetFPGA: Rapid prototyping of networking devices in open source," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 363–364, 2015.

[13] R. Ricart-Sanchez *et al.*, "Towards an FPGA-accelerated programmable data path for edge-to-core communications in 5G networks," *J. Netw. Comput. Appl.*, vol. 124, pp. 80–93, Dec. 2018.

[14] A. Huang, N. Nikaein, T. Stenbock, A. Ksentini, and C. Bonnet, "Low latency MEC framework for SDN-based LTE/LTE-A networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[15] N. Nikaein, X. Vasilakos, and A. Huang, "LL-MEC: Enabling low latency edge applications," in *Proc. IEEE 7th Int. Conf. Cloud Netw. (CloudNet)*, Tokyo, Japan, Oct. 2018, pp. 1–7.

[16] N. Mckeown *et al.*, "OpenFlow: Enabling innovation in campus networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, 2008.

[17] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A flexible and programmable platform for software-defined radio access networks," in *Proc. 12th Int. Conf. Emerg. Netw. Exp. Technol.*, 2016, pp. 427–441.

[18] *Docker*. Accessed: Oct. 2018. [Online]. Available: https://www.docker.com/

[19] *Kubernetes*. Accessed: Oct. 2018. [Online]. Available: www.kubernetes.io

[20] S. Compiler. (2017). *Xilinx*. [Online]. Available: www.xilinx.com

[21] (Apr. 2016). *Xilinx*. [Online]. Available: www.xilinx.com

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