SDN-based UPF for Mobile Backhaul Network Slicing

Jose Costa-Requena, Aapo Poutanen
Department of Communications and Networking
Aalto University, Finland

Serdar Vural, George Kamel, Chris Clark
5G Innovation Centre, University of Surrey
Surrey, UK

Shourov Kumar Roy
Advance Mobile Technologies Department, Cumucore Oy,
Finland

***Abstract*— A major challenge in future mobile networks is how to fulfil the requirements set for 5G networks in terms of latency and throughput. 3GPP has defined a new architecture based on virtualization and Software Defined Networks (SDN) supporting network slices that can fulfil those requirements. In this paper, we present the first realization of the new 5G user plane function (UPF) component that supports SDN and provides optimized transport for reducing latency as required in 5G networks. The proposed UPF is the cornerstone for using different data transport strategies adapted to the needs of different verticals, such as Ultra Reliable Low Latency Communications (URLLC) services that require a separate network slice providing an optimized transport for URLLC applications. The paper also discusses how to best migrate from legacy 4G user plane to 5G UPF, so as to ensure a reasonable transition to 5G. In doing so, the objective is not to meet short-term needs but to fulfill the future latency and throughput requirements of emerging applications, and also to provide first performance results of UPF in a realistic testbed environment.**

Keywords—5GaaS, UPF, Network Slicing

I. INTRODUCTION

5G has set ambitious requirements in terms of latency and bandwidth on mobile networks. In the coming years, traffic demand and the variety in services supported by mobile networks are expected to increase, not only with the usage of high resolution video (4K, 8K video streaming) but also due to the higher number of connected IoT devices. The current mobile networks rely on IP/GTP network shaping. However, the current protocol stack based on IP/GTP is not effective for avoiding network congestion or at least isolating the congestion to ensure reliability. In particular, the traffic prioritization at the IP layer is not sufficient to guarantee the level of reliability required by industrial or mission critical applications. Thus, the IP/GTP traffic shaping is inadequate and might not work properly when the network is congested.

Applications with low latency requirements require guaranteed service provisioning regardless of the network congestion and even under unexpected traffic load. Thus, a straightforward solution would be to reserve a physical link connection (OSI Layer 1), or to reserve links at Layer 2. This means that logical links at L1/L2 are reserved for the traffic flows that require low latency. However, this approach does not lead to an efficient utilization of available network resources. The L1/L2 link reservation will result in having idle resources when the traffic sent is below the allocated capacity. A solution in which L2 links are statically allocated to the services that require low latency (without having any constant demand) leads to resource under-utilization. Therefore, a dynamic allocation of L2 links is a desirable functionality for a mobile network. Software Defined Networking (SDN) provides the capability to dynamically change traffic flow descriptions, via various approaches, such as modifying the number of L2 links based on the number of traffic flows that require low latency. This approach allows a traffic flow to continue to receive the best possible service at the GTP/IP layer, but through SDN dynamic allocation of L1/L2 links will guarantee the reliability and low latency for certain flows.

In this paper, we present an SDN-based approach to the user plane traffic flows between the Radio Access Network (RAN) and the mobile core. The L2 links are realized either using physical or virtual ports in a physical or virtual switch that implements a programmable 5G User Plane Function (UPF). We have deployed L2 links as virtual connections using Ethernet VLANs. The user plane traffic flows coming from the base stations over GTP/IP are terminated in the UPF which then proxies the traffic to the next switch using dedicated VLANs that resemble the reserved L2 links required by low latency applications. The UPF can then apply different actions or priority to the traffic flows on those VLANs based on the latency requirements. SDN concepts are used to set traffic flow rules into the UPF switch, to support dynamic traffic differentiation at the user plane.

The result of the proposed approach is that network slices based on dedicated L1/L2 links can be implemented using the UPF that replaces GTP/IP with VLAN tunnels. SDN is used to enforce fine-grained traffic management in the VLANs compared to L3/L4 traffic throttling at IP/GTP layer. This solution creates the basis for 4G/5G network slicing.

In this work, we have deployed the first release of 3GPP-defined UPF in two different testbed environments in order to collect measurements from different infrastructures. The first

measurements are obtained from the testbed in 5G Innovation Centre (5GIC) [1] in University of Surrey, and are conducted as part of EU SoftFIRE project [2]. The second set of measurements are obtained from the testbed in AALTO University [3] as part of the TAKE5 project [4].

The results from both deployments show the benefit of deploying UPF as a stand-alone component performing local breakout for user data plane required for Mobile Edge Computing (MEC) in mobile networks. Performance results show that UPF allows effective implementation of network slicing which is highly desirable to support URLLC communications. The rest of the paper is structured as follows. Section II describes the 4G and 5G architectures and presents the issues with current deployments. Then, Section III introduces the proposed solution to overcome the limitations encountered with the 4G architecture when using virtualization techniques. Section IV presents the testbeds that are used to collect the measurements from deploying the proposed solution in a real infrastructure. Finally, conclusions are presented in Section V.

II. 5G MOBILE NETWORK ARCHITECTURE AND SDN OVERVIEW

The 4G/LTE architecture already separates control plane from user plane as shown in the upper part of Figure 1, where signaling is handled by the Mobility Management Entity (MME) and user plane is handled by the Serving and Packet Gateway (S/PGW). 5G architecture proposes a deeper separation between signaling and user plane by defining even smaller components within the control and user plane network elements as shown in lower part of Figure 1.

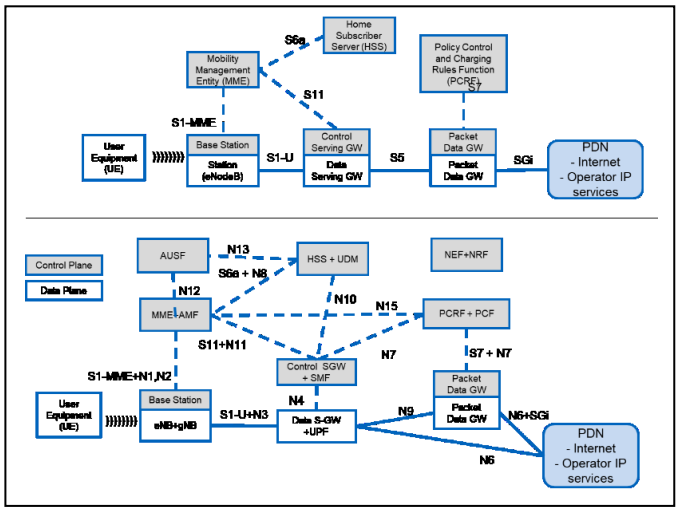

Fig. 1. 4G and 5G system architectures

The MME is now separated into Access and Mobility Management Function (AMF) and the Session Management Function (SMF). The user plane is managed together by the SMF and the User Plane Function (UPF). In addition to the decomposition into smaller network elements, the 5G architecture is proposing a new service-based architecture where the network elements are registered and discovered through a system repository named Network Repository Function (NRF).

The 4G architecture allows to virtualize each and every network component, and defines Virtual Network Functions (VNF) that consist of the virtualized network elements such as MME and S/PGW. Virtualization has a significant impact on mobile networks, as new elements such as network orchestrator or management network orchestrator has now been defined. There are several standardization groups both in SDN (Open Networking Foundation – ONF [5]), mobile networking (ETSI Management and Orchestration (MANO) [6]) contributing to the specifications of virtualization guidelines to be used in mobile networks. Moreover, a new set of open source tools and components (e.g Openstack) have been released lately to contribute to the adoption of virtualization.

The benefits of virtualization in mobile networks are significant. Currently, virtualization mainly facilitates scalability and automated management of network resources in data centers. In 4G/5G networks, virtualization allows to automate the launch and scaling up or down of virtual network functions based on traffic demand. The orchestration of mobile network elements in a data center allows to run all the required mobile network functions in virtual machines (VM) and increase or reduce the number of those VMs according to the traffic load. Figure 2 shows the usage of Network Functions Virtualization (NFV) [7] to scale S/PGW capacity, where new VM can be launched to run S/PGW nodes in the data center. The S/PGW will handle the traffic peaks and they will be terminated as soon as the traffic demand decreases, thus releasing the extra VMs and the allocated resources in the data center.

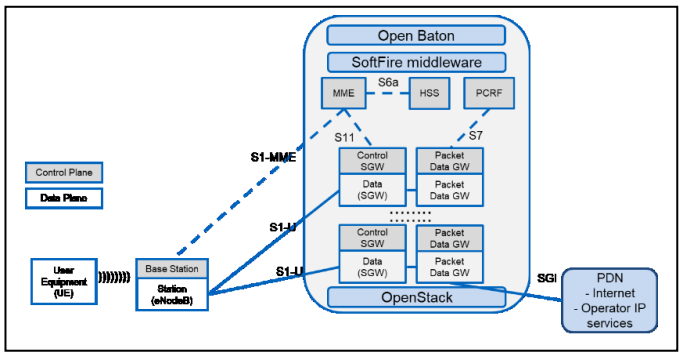

Fig. 2. Scaling of S/PGW as a virtual network function

Figure 2 describes the standard scenario in which virtualization is utilized for both control plane and user plane; the benefits reside in easy management of mobile functions. Mobile components are managed as usual software elements, which enables use of off-the-shelf software virtualization components.

The network can virtualized as a whole, i.e. either as a single virtual network function (VNF) or collection of VNFs, instantiated in the same data centre. This causes the user plane components to be potentially away from RAN, which is acceptable for signaling components but is not desirable for user plane that require MEC features. The virtualization of the user plane components means that all user plane traffic flows are terminated in the data centre, causing the traffic path to get congested. This deployment is not effective when there is high demand for network services that can potentially stay local and

be served locally. To address this issue, this paper proposes to virtualize individual components of the mobile core network separately. This deployment allows to place UPF closer to the edge of the network, providing user plane services closer to mobiles. The UPF integrated with the SDN will terminate the IP/GTP plane and will output dedicated VLANs to guarantee reliability and low latency.

Figure 3 shows the impact of traffic congestion when applying different types on L3 Diffserv per hop behavior traffic priorities such as Best Effort (BE), Critical cat:5 (CS5), expedited forwarding (EF) and Background service (BK).

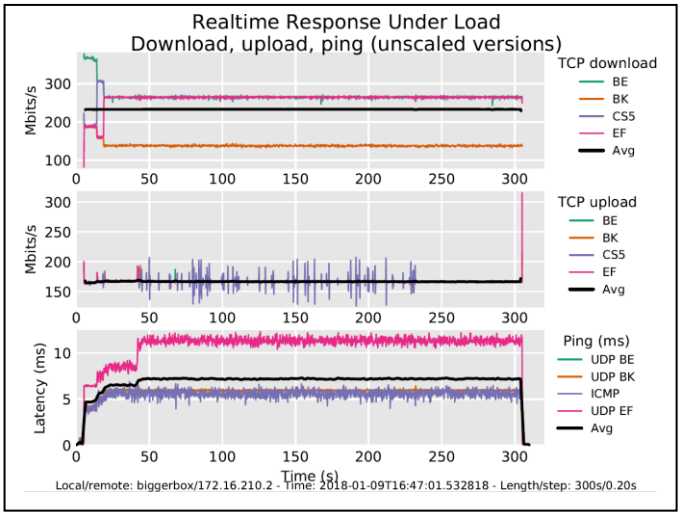

Fig. 3. User plane performance of UPF when deployed as a virtual network function in case of network congestion after applying different Diffserv traffic priorities (BE, BK, EF, CS5).

As seen in the Figure 3, in case of congestion even ICMP packets can be delayed multiple milliseconds on a direct 1GbE link even with lowest possible driver queue of 80. While effective to an extent, IP/GTP traffic classification alone does not provide significant improvements on latency in a congested environment. In order to provide guaranteed service level of below 1ms latency a L2 switched environment is required.

To achieve guaranteed service level a slice of user data network resources is allocated for critical applications. The proposed UPF node is placed on a point in the network to remove unnecessary tunneling as early as possible. The UFP terminates the IP/GTP tunnels and introduce the traffic classes into dedicated L1/L2 links to be propagated through the switched Ethernet network. The critical traffic is then routed through a non-congested link with allocated resources to ensure low latency. The traffic that requires low latency is inserted into dedicated L1/L2 links and they are differentiated by the information received from the AMF based on the mobile device SIM card.

When the UPF is co-located with the AMF and SMF on the cloud, it is observed that L3/L4 traffic classification is inefficient and cannot prevent the diverse effect of mobile backhaul congestion on user plane throughput and latency. On the other hand, deploying multiple VMs on the cloud, each running a separate instance of UPF can alleviate the scalability problem, as this provides easy and dynamic allocation of network resources.

We utilize UPF on the cloud to map the incoming GTP/IP traffic into IP traffic and utilize L3 prioritization based on the IP type of service (ToS) field according to DiffServ traffic shaping using DS, i.e. the DSCP 6-bit field. We observe that if the link between the RAN and the UPF is congested, the performance in terms of bandwidth and latency is heavily affected, despite the IP priorities we assign to user plane traffic. Therefore, we propose further separation of the user plane into modules that are managed by the control plane which can reside in the cloud. This approach is in line with the 5G architecture where the packet core can have multiple instances of UPF in different locations of the network. UPF can be deployed in any part of the network including the cloud. The capacity of UPF can be configured based on available data transport resources in the backhaul. The UPF can be deployed in a SDN switch but if existing switches cannot host a UPF, then a separate Linux based server can run the UPF. If the SDN switch has sufficient resources in addition to the UPF it can perform MEC application processing close to the RAN.

III. USER PLANE FUNCTION SDN INTEGRATED

In order to meet the demanding requirements of 5G networks this work proposes an efficient separation of the UPF. The 5G architecture is setting new possibilities to optimize the user plane by separating network elements that handle the user plane. Moreover, 5G also supports the addition of new transport technologies in addition to IP, thus Ethernet and so called non-structured transport is supported. In this paper, we propose the deployment of UPF as a separate programmable module integrated with SDN switch. Figure 4 presents the proposed implementation of UPF and its relationship with the 5G network architecture.

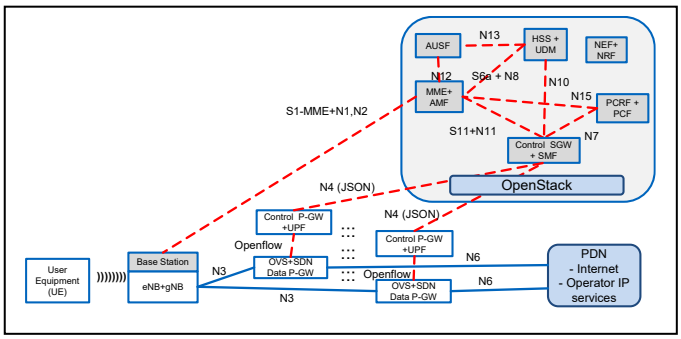

Fig. 4. 5G architecture with scaling of UPF fully distributed across the mobile backhaul while AMF and SMF are deployed on the cloud.

The major impact of the proposed solution is the improved granularity of control that EPC can have on the user plane traffic flows. This modular UPF enables the possibility of having network slices with different requirements in terms of latency, bandwidth and type of transport used. Thus, different UPF modules can be assigned to different slices or each user can have their own UPF to provide a different set of transport functionalities. This is necessary when the same network has to accommodate traffic over IP from end user services as well as non-IP traffic from IoT applications.

Despite its clear benefits, having a large set of UPF modules scattered over the mobile backhaul might lead into orchestration or management problems.

IV. SOFTFIRE AND TAKE5 EXPERIMENTATION SETUP

The proposed solution has been implemented as software to be deployed on virtualization platforms (i.e. OpenStack in this case) and performance evaluations presented in this paper aim to demonstrate its efficiency in realistic environments. In the current prototype solution, different users or applications are allocated to network slices based on the International Mobile Subscriber Identity (IMSI) numbers. This solution applies SDN to interact with the UPF module adding flexibility to apply different tunneling solutions on the user plane. The 5G AMF functionality that replaces the 4G MME, and the 5G SMF that has an S11-like signaling are implemented as Virtual Network Functions (VNF). The UPF handles legacy GTP tunnels and behaves as a Serving and Packet Gateway (LTE S/PGW) for supporting the user plane traffic to/from off-the-shelf LTE base station equipment. As replacement for SGW, the UPF connects to a switched L2 network of a network slice directly after the radio interface.

We have tested the system in two different testbeds. The first one took place in 5G Innovation Centre (5GIC) [1] in University of Surrey, and the setup in Figure 5.

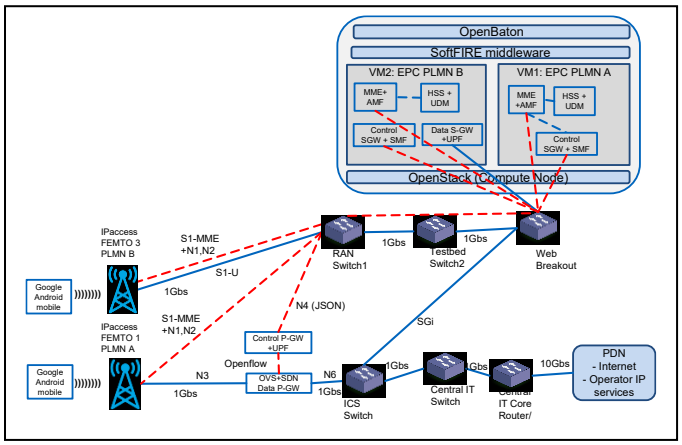

Fig. 5. Testbed 1 deployed with OpenBaton and SoftFIRE middleware in 5G Innovation Center, University of Surrey.

In this first setup, the SDN switch is placed in the mobile backhaul. The SDN switch is connected to the same network as the eNBs and obtains its IP address from the same DHCP server. The UPF running in the switch acquires the IP address in the same way but it is pre-configured based on its MAC address in the DHCP server so that the same IP address is always assigned to the UPF. The SDN switch also acquires a second IP address from the Campus network of the University of Surrey which is used to deliver the traffic from mobiles towards the public Internet.

The EPC is deployed in Openstack via the SoftFIRE middleware which communicates with the OpenBaton orchestrator [8] to deploy VMs at the 5GIC component testbed Virtual Infrastructure Manager (VIM) OpenStack, in SoftFIRE project. We deployed two instances of the EPC in Openstack and assigned each instance to a different operator or service. Each EPC instance is launched in separate VM and associated to a different mobile operator code (i.e PLMN A and B) as depicted in Figure 5.

Scenario 1: The UPF implementation (Data plane of SGW) is located in the same VM (VM2 Figure 5) where the control plane components of the EPC (i.e. MME, HSS, and the control plane of SGW) reside. This VM has connectivity to Femto 3 in the figure (femto-cell PLMN B). Both the user plane and the control plane of the mobile with PLMN B is connected to the EPC running in VM 2.

Scenario 2: In this scenario a separate VM (VM 1 Figure 5) which the control plane of the EPC (i.e. MME, HSS, and the control plane of SGW). The difference to Scenario 1 is that the user plane of the SGW is outside this VM, and has been implemented as UPF, running on an OVS based SDN switch, right next to the RAN (Femto 1, with PLMN A). This is a UPF deployment at the edge of the network.

The SDN switch receives the information from the EPC about the users that should get higher priority, lower latency or higher bandwidth. The EPC uses the mobile IMSI numbers to inform the SDN switch about selected mobiles and assign them different priorities in the mobile core user plane network. In the 5GIC testbed, we collected measurements about delay and bandwidth directly from the mobile devices connected to different base stations. Figure 6 shows the results after the installation, for measuring bandwidth and delay from Android mobile devices.

The results show the optimal bandwidth achieved when the installation reaches the best available performance values provided by the existing infrastructure i.e. 40,18Mbps downlink throughput, 15,71Mbps uplink throughput, and 42ms round trip time (RTT) delay.

Fig. 6. Measurements of bandwidth and latency using the Speedtest application on Google Android devices connected to IPaccess eNB equipment, as described in the setup of Testbed 1 at 5GIC Surrey.

Figure 7 shows the latency comparison results between the two scenarios; Scenario 1 with the vEPC slice where the EPC including both signaling and user plane (i.e. S/PGW-UPF) is entirely virtualized and running on the cloud. Scenario 2 with the SDN switch where the UPF is deployed at the edge of the RAN. No significant differences in terms of latency have been observed, since in both deployment scenarios, the edge UPF and the VM are essentially in the same data centre; i.e. the true

performance difference of MEC implementations are noticeable when UPF is distributed and deployed at remote locations, where eNodeB equipment is away from the mobile core data centre.

The experiment shows that it is technically possible to separate the UPF and deploy it at the edge of the RAN, without sacrificing performance or disrupting user plane operations. Another expected benefit of having an optimized SDN-based UPF arises when the network gets congested so the user plane traffic can be offloaded as an SDN local breakout, instead of routing all the user plane traffic up to the cloud.

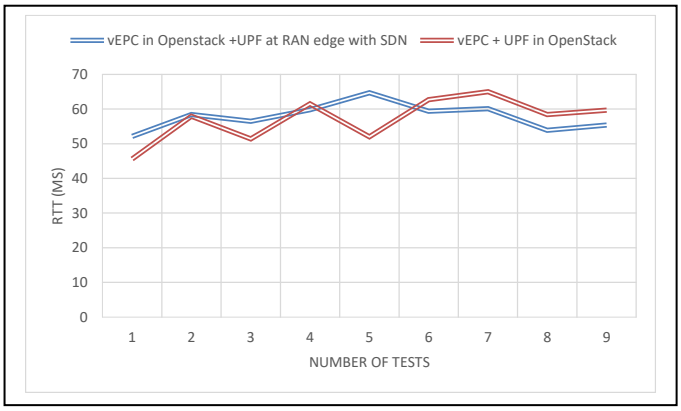

Fig. 7. Latency comparison between Scenario 1 (vEPC and UPF in OpenStack) and Scenario 2 (UPF at RAN edge at OVS SDN switch).

The second set of measurements were taken in Testbed 2 deployed by the TAKE5 project [4] in the campus of AALTO University [3]. This testbed consists of real deployment of 10+ Nokia macro-cell eNBs and 20+ Nokia indoors pico-cell eNBs connected to different EPCs. The testbed includes a Nokia commercial EPC and two other experimental EPCs, one running on Openstack and another running on a dedicated Linux server. Figure 8 shows the setup in the Testbed 2 similar to Testbed 1 with two scenarios; Scenario 1: Both signaling (MME, HSS, Control SGW) and user plane (S/PGW-UPF) are deployed in the cloud on Openstack.

Scenario 2: The signaling (MME, HSS, Control SGW) is on the cloud on OpenStack, and the user plane (S/PGW-UPF) is deployed in a separate server located at the edge of the RAN. The UPF terminates the GTP/IP tunnels and output the UE IP traffic into L2 tunnels.

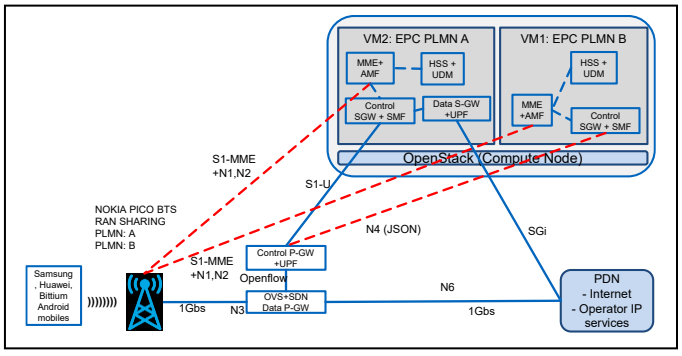

Fig. 8. Testbed 2 with Openstack and UPF in mobile backhaul deployed in AALTO University part of TAKE5 project setup.

The difference between the two testbed deployments is the equipment used in the RAN; Testbed 2 has Nokia eNBs. This RAN equipment provides a different level of throughput and latency on the air link. In Testbed 2, we also use a larger number of device models (i.e. Huawei, Nokia, Bittium and Samsung). Moreover, in this second testbed, we explicitly apply some level of congestion to the user plane link between the RAN and the cloud (S1-U) to evaluate the impact of deploying UPF in the mobile backhaul instead of the cloud.

The impact on user plane throughput in Testbed 2 (with the congested link) is similar to the results shown in Figure 6. However, the major impact of deploying UPF at the RAN edge (to avoid the link congestion on the S1-U/N3 interface) is on the observed user plane latency, as we can see in Figure 9. The figure demonstrates the difference in latency between the two scenarios; i.e. Scenario 1 where UPF does not apply any tunneling and runs on OpenStack, and Scenario 2 with UPF deployed on an SDN switch in a dedicated slice, and controlled by the EPC in OpenStack via L2 tunneling. In Figure 9 we can observe how the SDN based UPF(Scenario 2) can map the GTP/IP into a dedicated L2 tunnel while maintaining a much more stable user plane latency performance over time.

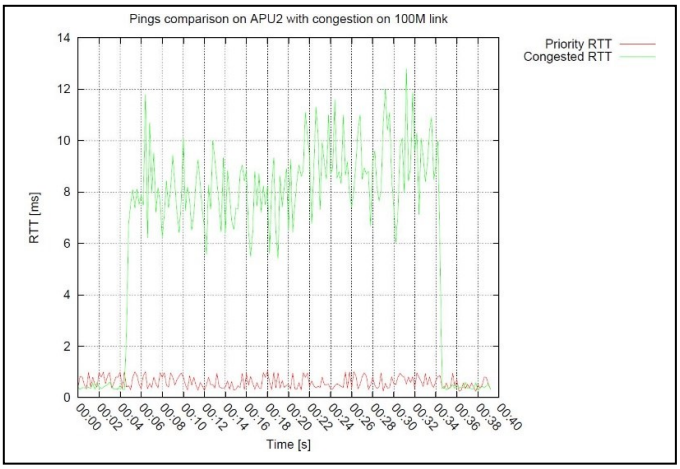

Fig. 9. Latency performance difference in congested link with UPF deployed at the edge of the RAN, which also applies traffic prioritization.

However, in Scenario 1, UPF does not provide any specific L2 tunnel, and suffers from the increase in latency when the network gets congested. The UPF is optimized for data delivery and current implementation has been tested in high bandwidth conditions to confirm that latency remains with the required limits. Figure 10 shows the UPF in the slice connected at the edge of RAN network so it keeps delivering stable latency of about 0,2 ms with peaks traffic of 20Gbs download and 14Gbs upload.

The results from Figure 10 are obtained using the UPF data plane running in an HP Z240 workstation with the 40Gbs NIC Intel [9] with QSFP transceiver and 1m DAC cable connected to the RAN. This deployment is applicable when low latency and high bandwidth are required for Multi-Access Edge Computing (MEC) applications. 5G networks require higher capacity backplane in the order of Tera bits (Tbs) to support the requirement of 1Gbs per user.

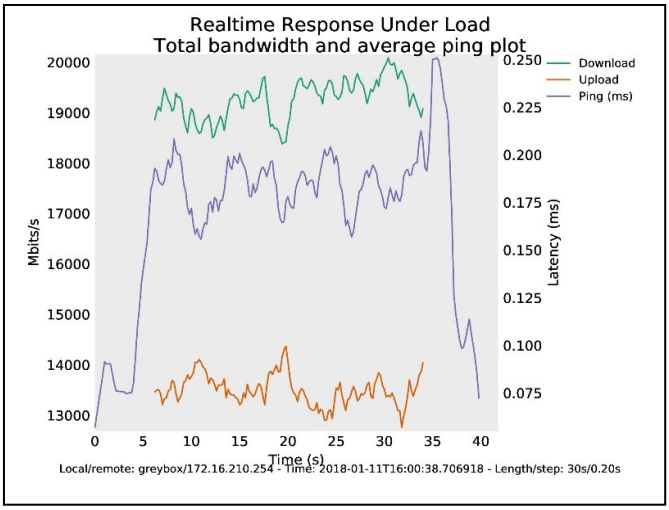

Fig. 10. Latency performance for high badnwidth requirements with UPF deployed at the edge of the RAN.

However, the usage of the proposed network slicing with the SW based UPF provides the required flexibility to handle the traffic on the RAN network. Thus, alleviating the backhaul and core network from higher traffic rates. Moreover, the fact of having a SW based UPF managed through SDN allows to dynamically configure or even migrate the UPF across the RAN network based on the number of users or applications that require the low latency and high bandwidth. The usage of UPF in commodity HW eases the deployment of network slicing together with MEC using SDN as technology enabler.

V. MEASSUREMENTS ANALYSIS

The testbed 1 in 5GIC we deployed 5G network elements using full orchestrated tools such as SofFIRE middleware and OpenBaton with Openstack. The results in Figure 6 and 7 show the latency and bandwidth available after deploying the 5G network fully virtualized. The testbed 2 in TAKE5 5G network elements were deployment similarly but in addition we congested with up to 40Gbs traffic as shown in Figure 10. We demonstrate the efficiency of using UPF for network slicing based on SDN and replacing GTP tunnels with VLANs. The benefits of UPF in the network stack as seen in Figure 11 where UPF is located in the eNB. UPF enables the usage of optimized transport technologies (i.e. VLANs, MPLS, GREs) in the mobile backhaul right after the eNB. This provides a dedicated network slice for selected traffic where prioritization is done at the lower layer of the stack.

In both testbeds UPF was mapping GTP into VLANs but as required by 5G standards the UPF can be configured to utilize other transports like non-IP which might be required for Machine to Machine (M2M) communications.

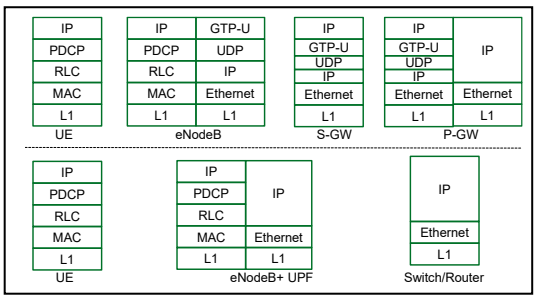

Fig. 11. User plane network stack with the UPF.

VI. CONCLUSIONS

This work presents the first deployment of the 5G architecture in real testbeds with a disruptive approach based on SDN based UPF. The proposed solution not only provides an incremental evolution from 4G towards 5G, but also a lower cost of implementing UPF based network slicing compared to a monolithic EPC as in LTE. The UPF integrated in a SDN switch provides network slicing by removing IP/GTP and using instead dedicated L1/L2 links to create the slices. The paper shows the results from the fully distributed UPF/SDN network slicing which is building block for URRLC and MEC. The modular SDN based user plane allows dynamic allocation of UPF in any part of the network as well as optimal management and orchestration of user plane resources. The system can scale up horizontally with additional UPF modules when traffic demand increases or vertically with additional instances of the 5G EPC. The 5G UPF facilitates the possibility of having many instances as needed each deployed in different part of the network to deliver MEC and each of them with different capacity and reliability to deliver URLLC.

ACKNOWLEDGMENT

This work was accomplished as part of a SoftFIRE experiment in 5G Innovation Center at University of Surrey, and in cooperation with AALTO University as part of the TAKE5 research project funded by the Finnish Funding Agency for Technology Innovation and industry.

REFERENCES

[1] 5G Innovation Centre, University of Surrey, https://www.surrey.ac.uk/5gic

[2] EU SoftFIRE project, https://www.softfire.eu/

[3] Department of Communications and Networking, Aalto University, Finland, http://comnet.aalto.fi/en/

[4] Take5 Project, http://5gtnf.fi/projects/take-5/. www.take-5g.org

[5] Open Networking Foundation (ONF), https://www.opennetworking.org/

[6] ETSI NFV Management and Orchestration (MANO)

[7] “Network Functions Virtualisation— Introductory White Paper”, ETSI, 22 October 2012, retrieved 20 June 2013.

[8] OpenBaton, https://openbaton.github.io/

[9] Intel 40Gbs (Ethernet Converged Network Adapter XL710-QDA2).