The authors introduce a lightweight, ETSI-compliant MEC solution for 4G and 5G networks. The proposed solution, which they name LightEdge, has the main goal of immediately making available the features and capabilities of edge clouds to mobile users. They report on the design and implementation of LightEdge and on its evaluation in a practical latency-sensitive use case.

**INTRODUCTION**

The fifth generation (5G) is opening the door to a new generation of applications and services. Augmented reality and holographic interfaces are but a few of the applications that will benefit from the massive bit rates and ultra-low-latency features expected to be provided by 5G [1]. For the support of ultra-low-latency services, multi-access edge computing (MEC) will be a key component in the 5G system architecture [2]. The 5G service-based architecture introduces the user plane function (UPF) concept, decentralizing the data forwarding functions from the control plane functions [3]. The UPF serves as the anchor point for mobile terminals and provides interfaces for traffic billing and lawful interception. It also enables application detection using service data flow filter templates or 3-tuple (protocol, server-side IP address, and port number) packet flow descriptions received from the session management function. This allows packet processing and traffic aggregation to be performed closer to the network edge, hence increasing bandwidth efficiency while remarkably reducing backhaul network utilization and service latency. This fact makes the integration and deployment of the MEC system in the 3rd Generation Partnership Project (3GPP) 5G architecture a convenient option.

Despite the benefits brought by 5G, it will take some time until the existing 4G system transitions to a full 5G system. Commercial 5G networks are initially being deployed in non-standalone mode, meaning that the 5G radio access network (RAN) still interfaces with the 4G Evolved Packet Core (EPC). While this approach immediately makes higher bit rates available, it impedes the deployment of new applications and services at the network edges. This is because the standard EPC lacks proper support for edge applications and services that are instead a native part of the 5G core. The challenge is thus to develop a transition mechanism allowing the integration of MEC systems in the 4G architecture to make their features and capabilities directly available to mobile users.

Several works have proved that it is possible to introduce MEC and edge-steering solutions in 4G networks [4–7]. However, none of them tackle the challenges and practicalities associated with the deployment of such solutions. In this article we propose LightEdge, a lightweight, European Telecommunications Standards Institute (ETSI)-compliant MEC solution for 4G and 5G networks. The main goal of LightEdge is to provide mobile network operators (MNOs) with an MEC platform that can immediately bring the advantages of edge computing to the 4G end users, while enabling a seamless transition over the evolutionary path from 4G toward a full 5G architecture.

LightEdge is designed to integrate seamlessly with standard ETSI network function virtualization (NFV) solutions like open source management and operation (MANO) (OSS) [8] and potentially leverage the features of the MANO as a service (MANOaaS) paradigm [9]. Although this work aims to propose a transparent MEC solution that enables some 5G features in a 4G network, the platform can also be used in a full 5G network operating in either non-standalone or standalone mode. In the latter case (standalone), one of the components on the LightEdge platform, that is, the UPF service, can be replaced by the standard UPF service already present in the 5G architecture.

This article reports on the design, implementation and evaluation of LightEdge. The platform has been validated with commercial and open-source eNodeBs (eNBs) and EPCs, and has been showcased by a MEC-enhanced, latency-sensitive application for driving assistance. Experimental results prove the seamless integration of LightEdge with the existing 4G and 5G architectures.
**Design: Requirements and Challenges**

The line between traditional core, transport, radio, and IT groups at current MNOs is becoming more blurred with changes and upgrades in one sector percolating into the others. This leads to severe organizational struggles when MEC offerings are to be integrated into the production networks. Therefore, the design of LightEdge has been driven by the following requirements:

- To minimize the changes to the MNO’s environment, with the possible exception of the online charging system and the offline charging system.
- To provide a platform that can be extended and evolved toward the 5G architecture.
- To comply with recent cloud-native trends allowing deployment of containerized edge applications.
- To provide a solution capable of supporting local breakout for enterprise applications.

Before explaining how LightEdge satisfies these requirements, it is important to recap the main components of the 3GPP 4G architecture (Fig. 1) to clearly identify the challenges it poses [10]. User equipments (UEs) connect to the RAN through the eNBs. The EPC is composed of the packet data network gateway (PGW), which acts as contact point between the serving gateway (SGW) and another data network (typically the Internet), the SGW, which is an intermediate aggregation point between the RAN and the PGW, the mobility management entity (MME), which is responsible for the handovers, and the home subscriber system (HSS), which acts as a database and contains the subscribers’ information. These entities are interconnected via well-defined interfaces.

The most relevant point with respect to LightEdge is the S1 interface, which defines the data/control plane protocol between eNB and SGW, and the control plane protocol between eNB and MME. The protocol between the eNB and the SGW, carried over UDP, is the GTP Tunneling Protocol (GTP). Conversely, the protocol between the eNB and MME is the S1AP protocol, which is carried over Stream Control Transmission Protocol (SCTP). The connection between a UE and the SGW is identified by the tunnel endpoint IDs (TEIDs). It is the duty of the MME to assign the TEIDs between UEs and SGW (upstream) and SGW and UEs (downstream). TEIDs are dedicated fields of the GTP-u message used to route the GTP tunnels. The traffic from/to a UE is encapsulated in a GTP tunnel, which includes an IPv4 header, a UDP header, and the GTP header (which includes the TEID). This encapsulation for the downstream traffic is also depicted in Fig. 1.

All data connections of a UE are anchored to a certain PGW. This means that even if the UE moves, its IP address to the Internet does not change. However, it also implies that if the UE is roaming in a foreign network, all its traffic is home-routed via its PGW. For example, an Italian UE roaming in Germany will see its traffic first routed to its home PGW in Italy and then to its final destination, resulting in considerable latency and high load on the backhaul and core links. In this scenario, introducing a MEC host in close proximity to the RAN allows terminating the UE traffic before reaching its PGW and serving it with the required services from the intermediate MEC host. Hence, it reduces the latency and the traffic load at both the PGW and the backhaul/core links. It is with this objective that we propose LightEdge as a 5G MEC solution to be integrated in the existing 4G systems.

**LightEdge System Architecture: An Overview**

The LightEdge system architecture, depicted in Fig. 2, is designed to allow UEs to consume applications and services at the network’s edges. LightEdge follows the bump in the wire architecture proposed by ETSI [11], thereby placing the MEC host between the RAN and the EPC of the 4G system to enable the interception of UE requests. Before reaching the intended DNS server, requests from users are resolved to the virtual IP address of a local MEC application. Note that this approach will not work if secure alternatives to DNS (e.g., DNSSEC) are employed. When a UE requests the virtual IP address of a local MEC application, LightEdge takes over the communication, and the UE traffic is steered toward the MEC host. Standard stateful L4 or L7 load balancers (not depicted in the figure to improve readability) can be used to distribute the load among multiple MEC application instances.

This approach allows minimizing the changes to the MNO’s existing infrastructure, where our proposal can be seamlessly deployed between RAN and EPC as long as access to the S1AP and the S1 interfaces is provided. Re-provisioning of MME pools is also avoided by intercepting relevant UE attachment, detachment, and mobility events using an S1AP Monitor module. This essentially translates into the ability to dynamically track the mapping between UEs and their TEIDs for multiple bearers. This information is then used by the vGTP module to reconstruct the GTP tunnel between MEC hosts and eNBs/gNBs. The rest of this section describes the LightEdge’s components and the mobility management and billing procedures in more detail.

**MEC Platform Extensions**

The design of LightEdge extends the ETSI reference architecture [12] and encompasses the functionalities required to run MEC applications, and...
to allow them to provide and consume MEC services. The LightEdge MEC platform is configured by a MEC Platform Manager via the Mm5 interface or by other applications and services via the Mp1 interface. Note that the MEC Platform Manager is outside the scope of this article in that our goal is to focus on the definition of a transparent MEC platform. Nevertheless, the MEC platform itself exposes a standard REST interface that can be used by present operations support system (OSS) platforms. Note that the Mm5 interface is not yet defined by ETSI, and hence no standard-compliant MEC Platform Manager is currently available. The MEC platform is then responsible for implementing such configurations including, for example, traffic rules and DNS entries. Here follows a brief description of the LightEdge specific sub-components extending the MEC platform.

Service Registry: The Service Registry is an ETSI MEC functional component that contains the catalog of services and applications that can be spawned on the MEC platform. An example is the Radio Network Information (RNI) Service [13], which provides MEC applications with real-time information about the RAN (e.g., RSRP/RSRQ measurements).

Radio Network Information (RNI) Service: In the 5G architecture the MEC platform can obtain UE information by the northbound and southbound interfaces. Data through the southbound regards UE location and roaming state from the network exposure function in the 5G core, while data through the northbound data regards radio bearer statistics from the radio resource manager. In this work, LightEdge focuses on the RANI information toward the RNI service.

Message Queuing Telemetry Transport (MQTT) Service: The MQTT service forms the communication nexus between the various sub-components of LightEdge. It works following a publish-subscribe paradigm where services and applications can publish new information on a certain topic and subscribe to one or multiple topics. For example, a video streaming application could subscribe to the messages containing the RSRP/RSRQ measurements of a certain UE and adapt the video transcoding accordingly. The openness, light weight, and ease of implementation of this protocol make it an ideal option for environments where processing, memory, and bandwidth efficiency is mandatory.

Traffic Rule Manager: The Traffic Rule Manager is in charge of (re)configuring the L3 switch to route the traffic among applications/services and the 3GPP network. Traffic rules are issued by an external MEC Platform Manager and enforced by an L3 switch. This feature allows LightEdge to tap into the S1 and S1AP interfaces and to redirect them to the S1AP Monitor module, which is part of the 4G-UPF Service.

DNS Resolver: The DNS Resolver allows mapping UE requests to local IP addresses routable inside the MEC domain. DNS records are filled based on a configuration coming from the MEC platform manager or following an activation request from the MEC applications. Any DNS resolver can be used for this purpose. The relationship between a local IP address and one or more physical IP addresses is handled by the virtualization infrastructure manager (VIM), which preserves the virtual reference even if a MEC application is reallocated or shut down, making the process fully transparent for the UEs.

4G-UPF Service

This is the core service that is complemented and leveraged by the LightEdge service extensions in the MEC platform. A detailed view of this component is included in Fig 2. This service comprises the functional elements described below.

The L3 switch is in charge of steering the traffic between eNB/EPC and the MEC services under the control of the Traffic Rule Manager. The S1AP channel is steered to the S1AP Monitor module by matching the IP protocol type (SCTP is 0x84), while the GTP-u stream is steered to the vGTP module by matching the UDP port (2152 for GTP-u). The L3 switch can be either a hardware-accelerated switch (e.g., P4-based) or the software switch at the Linux kernel. The choice depends on the number of eNBs connected to the MEC host and on their configuration. For example, a 20 MHz LTE cell using a 2 x 2 multiple-input multiple-output (MIMO) configuration results in a maximum theoretical downstream bit rate of 150 Mb/s. Modern software switches can easily handle tens of such cells. Conversely, larger deployments and/or wider bandwidths may require hardware-accelerated switches.

The S1AP Monitor receives the SCTP-encapsulated S1AP traffic and tracks the upstream/downstream TEIDs. This is done by monitoring the InitialContextSetupRequest and the InitialContextSetup.
textSetupResponse messages. The former assigns the upstream TEID, while the latter assigns the downstream TEID. Once both messages are collected for a UE, a new rule is added to the UE/TEID Mapping Table (as shown in the 4G-UPF Service component in Fig. 2). It is worth noting that the messages exchanged using the S1AP protocol are typically encrypted. This issue can be overcome by passively sniffing the traffic on the S1-u interface in order to learn the UE/TEID mapping. This approach would, however, make LightEdge not suitable for mobility scenarios where sniffing the S1-u interface is not enough to properly detect handover events. Alternatively, we can envision a service-based architecture whereby a proxy implemented within the S1AP Monitor subscribes to mobility events published by the MME. This approach would allow keeping the UE/TEID Mapping Table up to date even in the case of handover events.

The vGTP manages the stateful GTP encapsulation/decapsulation between eNB and SGW. The inner IP flow of the upstream GTP-u traffic is matched against the Traffic Steering Table. If a match is found on the 3-tuple (protocol, server-side IP address, and port number), the GTP protocol stack headers (GTP, UDP, and IP) are removed.

If an entry is present in the Virtual IP/Port column of the Traffic Steering Table, the destination IP address and the transport port of the matched flow are rewritten to the values specified in the Traffic Steering Table. In the example in Fig. 2, flow (0x06, 10.224.52.113, 80) is left unchanged, while the destination IP address and TCP port of flow (0x06, 3.120.16.110, 80) are rewritten. Both flows are then handled by the standard Linux NAT/PAT subsystem. Notice that at this point, due to either the DNS/Remote Server redirection or the address/port rewriting performed by the vGTP, the destination address of the decapsulated IP flow is an address routable in the local MEC domain. IP traffic from a MEC application instance and addressed to a UE is encapsulated into a GTP tunnel by the vGTP using the UE/TEID Mapping Table.

**Handling User Mobility**

The MEC platform must be able to keep serving users even when handovers occur. In this work, we limit our attention to X2 handovers since they are the most common option used by operators to implement this procedure. Nevertheless, the proposed approach can also be extended to S1 handovers. This part is omitted due to space constraints. Figure 3 depicts the messages exchanged during an X2 handover.

A UE detecting that a neighboring cell has better channel quality can trigger an X2 handover. When this happens, the source eNB sends an X2 handover request to the target eNB. The target eNB then issues an S1AP Path Switch Request to the MME. Since the S1AP Monitor processes all S1AP signaling, it can identify the start of an X2 handover and the IP addresses of the eNBs involved. As a result of the path switch request, the MME establishes a new GTP tunnel between the SGW and the target eNB. After the X2 handover is acknowledged by the new eNB, the UE attaches to the new eNB.

The MME can then modify the path for all bearers of the UE and inform the SGW of this change using a Modify Bearer Request message. The handover is completed when the S1AP Path Switch Request ACK is sent by the MME to the target eNB. When this happens, the S1AP Monitor updates the corresponding TEIDs in the UE/TEID Mapping Table. Afterward, the S1AP Path Switch Request ACK downstream packets are forwarded to the new eNB using the newly established GTP tunnels. This tunnel is then intercepted again by the vGTP, which performs the stateful GTP encapsulation/decapsulation again.

**Traffic Charging and Billing**

In a 4G network the charging trigger function (CTF) is responsible for intercepting chargeable events, for example, data volumes, sessions start/stop, and handovers, and for building the charging data records (CDRs), which are then sent to the mobile network operator’s (MNO’s) billing system through the Diameter protocol. The CTF is implemented by the PGW. However, since in LightEdge the traffic to/from a MEC application does not traverse the PGW charging function, that traffic becomes impossible to track.

To solve this issue, we leverage a novel proposal in [14], which introduces a delegated chargeable event monitoring function (D-CEMF) responsible for capturing charging events at off-load points. For LightEdge, the D-CEMF is proposed to be deployed within the 4G-UPF service, where it can intercept the S1AP events and track the offloaded S1 using the per-UE packets and byte counters maintained by the vGTP. The CTF in the PGW can then aggregate information coming from multiple D-CEMFs to build the consolidated CDRs.

**Implementation Details**

LightEdge has been designed with cloud-native principles in mind. Each of its components is deployable using container technologies, and the platform itself is natively compatible with Kubernetes. MEC applications can be deployed as con-
containers and leverage the full capabilities of the container networking interface (CNI) used by modern cloud-native environments. Notice that our design is agnostic with respect to the particular CNI technology employed and can work even without a CNI. The only requirement is to have a switched L3 network. This last feature makes LightEdge particularly suitable to support local breakout for enterprise applications.

We have developed a prototype of LightEdge and deployed it on an LTE testbed. The RAN part comprises a 3GPP-compliant LTE stack provided by srsLTE, while as the EPC we use nextEPC. It must be noted that LightEdge is vendor-agnostic and can be used with any combination of eNB/EPC components (including commercial ones). The 4G-UPF service is implemented as a native application, while the MEC platform is implemented as a simple Python agent. The Traffic Rule Manager and the DNS Resolver are implemented using, respectively, iptables and dnsmasq. Following the control and user plane separation (CUPS) concept [10], the 5G-EmPOWER software-defined RAN controller [15] implements the radio resource manager. The control plane interface of this system is based on the ETSI RNIS MEC application programming interface (API) [13] and provides a two-fold function: RAN elements configuration and RAN-level statistics collection, which are then exposed to the RNI service. 5G-EmPOWER can interact with commercial and open source eNBs. Finally, the MQTT service is implemented using RabbitMQ.

**AUTONOMOUS DRIVING: A PRACTICAL USE CASE**

For the evaluation of LightEdge we have selected connected, cooperative, and automated mobility (CCAM) as a practical use case. By using state-of-the-art MEC solutions like LightEdge, autonomous driving functions can be offloaded to the network, thereby relieving vehicles from the computation burden. In the next sections we describe the general requirements of CCAM applications and discuss the evaluation on the LightEdge platform.

**CCAM APPLICATIONS REQUIREMENTS**

- **Lane Tracking**: This application processes in real time the images from cars to ensure safe maneuvering, and computes and sends back the corresponding steering commands. Note that this operation is computationally expensive and involves high uplink bandwidth consumption.

- **On-Road Object Recognition**: This application enables the detection of entities such as plates and signs based on computer vision classifiers. The computational and bandwidth requirements are analogous to the ones described above since video streams from vehicles also need to be analyzed.

**PERFORMANCE EVALUATION**

The evaluation aims to prove the ability of LightEdge to:

- Manage various MEC system configurations and distribute the traffic load across several MEC application instances in a transparent manner for the end user
- Work seamlessly on different RAN configurations and manage requests from several base stations, delivering the adequate latency level

The CCAM applications (i.e., assisted driving operations) have been deployed on the LightEdge platform. To the best of our knowledge, there is no other MEC platform transparent to any RAN and core network. Therefore, we have included in the evaluation a cloud setup (inherently transparent to RAN and core) to show the capabilities of LightEdge and not to merely draw a performance comparison. LightEdge is deployed on a machine analogous to the c4.xlarge Amazon EC2 equipped with an Intel Xeon E5-2666 processor with 4 vCPUs and 7.5 GB RAM memory, while the cloud setup leverages c4.4xlarge instances deployed in Ireland comprising 20 vCPUs and 30 GB RAM memory. The evaluation is performed on the experimental testbed described previously. The aforementioned objectives are illustrated in three scenarios with diverse MEC host and RAN configurations:

![Figure 4. Latency vs. an increasing number of requests for a single MEC application instance and a single RAN element.](image-url)

![Figure 5. Latency over time for an increasing number of requests (one every 2 s). An L7 balancer distributes the load to a second application instance when the latency exceeds a threshold.](image-url)
Single Application Instance — Single RAN Element: Figure 4 reports the average latency experienced for an increased number of requests received by a single application instance in a setup comprising one eNB. This latency includes the round-trip time (RTT) and the application computation time. In particular, it can be seen how the RTT is slightly affected by the computation burden. This is visible in LightEdge when reaching six requests. Although this issue applies to a lesser extent to the cloud deployment due to the greater computational resources, the RTT offered by LightEdge is still lower regardless of the number of requests. Moreover, although the cloud provides shorter computation time, the overall latency exceeds the time taken by LightEdge for request processing due to the backhaul load reduction.

Several Application Instances — Single RAN Element: The latency caused by the CPU load in Fig. 4 compromises fast-response services. To lower the computational intensity, we introduce into the previous scenario a stateful L7 balancer, which distributes the load among MEC application instances when the total latency exceeds 150 ms. Notice that, if not done automatically, the MEC Platform Manager must request that the VM horizontally scales the application server and deploy a new application instance before the load distribution takes place. Figure 5 shows the results over time of this topology, where a new request is generated every 2 s. Specifically, it can be seen that after 12 s the delay threshold of the balancer is exceeded. As a result, new incoming requests are dispatched to a new application instance, thus reducing the execution and access delay with respect to cloud computing.

Several Application Instances — Several RAN Elements: Taking as baseline the setup shown in Fig. 4 (given by one application instance and one eNB), Fig. 6 analyzes two configurations regarding the RAN and the MEC system to evaluate how both segments determine the latency experienced. The first setting maintains the RAN configuration and focuses on the MEC host. Two application instances are initially spawned in such a manner that requests are evenly distributed across them attending to the delivered latency. The second setting extends the first one by adding a second eNB. As can be seen, the last configuration achieves the best results as the number of requests increases given that greater computation capacity and radio resources are available. Nevertheless, it is important to highlight the low difference with respect to the configuration comprising a single eNB given that the computation delay at the edge represents the largest portion of the end-user latency.

CONCLUSIONS

Multi-access edge computing is a promising paradigm for future mobile networks. In this article, we introduce LightEdge, a lightweight, ETSI-compliant MEC solution for 4G and 5G networks. The design proposed is transparent to the existing 4G networks, therefore requiring almost no changes to the MNO’s environment. LightEdge is well suited to serve the needs of different verticals, including smart cities, augmented reality, and connected and cooperative road mobility. In particular, we have shown how LightEdge can be used to offload autonomous driving operations to the network’s edge.

These operations could be split into atomic tasks to be independently executed locally or remotely, for example, image transcoding from vehicles’ cameras could be performed locally before being processed remotely. This brings new challenges on operation decomposition into tasks, execution site selection, and task synchronization for optimal performance. In addition to such challenges, we plan to enhance the interplay between RAN and LightEdge, and to embed the logic to enable autonomic scaling of MEC resources.

ACKNOWLEDGMENTS

This work has been performed in the framework of the European Union’s Horizon 2020 project 5G-CARMEN co-funded by the EU under grant agreement No 825012. The views expressed are those of the authors and do not necessarily represent the project. The Commission is not liable for any use that may be made of any of the information contained therein.

REFERENCES


**Biographies**

**Estefanía Coronado** (e.coronado@fbk.eu) is an expert researcher in the Smart Networks and Services Unit at FBK. In 2018 she completed her Ph.D. at the University of Castilla-La Mancha, Spain, on multimedia distribution over SD-WLANs using ML. She received M.Sc. degrees in computer engineering and advanced computer technologies in 2014 and 2015 from the same university. She has published around 20 papers in international journals and conferences, and received the IEEE INFOCOM Best Demo Award 2019. Her research interests include wireless and mobile communications, network slicing, SDN, NFV, AI-driven networks, and automated network management.

**Zarrar Yousaf** is a senior researcher at NEC Laboratories Europe, Germany. He completed his Ph.D. from TU Dortmund, Germany. His current research interest is in NFV/SDN in the context of 5G networks. He is also an active contributor to the ETSI ISG NFV standards organization where he holds Rapporteurship for six work-items. He has extensive experience in the design, development, modeling, simulation, and prototyping of communication systems and protocols for optimizing overall network performance. He has four granted patents and 19 filed patents, while his research work has been published in several peer reviewed journals, conferences, and book chapters.

**Roberto Riggio** [SM] is head of the Smart Networks and Services Unit at FBK. His research interests include smart networks and services, network slicing, and distributed management and orchestration of network services. He has published more than 120 papers in journals and conferences, and has generated more than €3 million in competitive funding. He has received several awards including the IEEE INFOCOM Best Demo Award (2013, 2017) and the IEEE CNSM Best Paper Award (2015). He serves on the TPC/OC of leading conferences in networking and is an Associate Editor for Wiley’s *International Journal of Network Management*, Springer’s *Wireless Networks* journal, and *IEEE Transactions on Network and Service Management*. He is a member of ACM.