2016 IEEE 1st International Workshops on Foundations and Applications of Self-* Systems

Enabling low latency services in standard LTE networks

Cesar A. Garcia-Perez, Pedro Merino
Universidad de Málaga, Andalucia Tech, Spain
{cgarcia,pedro}@lcc.uma.es

***Abstract*—The upcoming 5G technologies promise to enable ultra low latency services such as remote robotics, augmented reality or vehicle to vehicle communications. Fog and MEC computing can enable low latency services for many different scenarios by moving the cloud and some network functions closer to the user. In the case of mobile networks this improvement is traditionally introduced in the reference point that defines the limit of the operator domain. In this paper we explore the introduction of an intermediate component in the LTE standard architecture, describing its functionality and providing experimental results on its use. This component, called Fog Gateway, can process the data plane for specific services to prevent all the traffic reaching the core network. The gateway analyzes the GTP traffic inner destination IP in order to determine whether to route the packet to the fog network or forward it to the destination SGW. The solution is compliant with standard LTE equipment all along the path (UE, eNodeB, EPC), so it can be implemented in current networks, and the preliminary figures, obtained combining emulated and COTS equipment, show an improvement of up to the 78% in terms of latency reduction.**

I. Introduction

One key objective in 5G networks is to reduce end-to-end latency so as to offer advanced critical services, like telesurgery, vehicle to vehicle communication or services in the *tactile internet* [1]. Gaining a few milliseconds is not possible without some changes in the current architecture of mobile networks. Fortunately, there are new possibilities thanks to the integration of more intelligent (software) in the network.

The Radio Access Network (RAN) will offer additional computation capabilities apart from the pure standard processing of the protocol stacks. At this edge of the network, it will be possible to deploy part of the services usually provided in the cloud, leading to the *FOG computing* paradigm [2]. The mobile network operator (MNO) can also use the new computing platform to replicate or move networks functions closer to the user thanks to *Mobile Edge Computing* (MEC)[3], [4]. Therefore the edges of the network can operate in an isolated environment with respect to the centralized components (like the Evolved Packet Core (EPC)).

At the core of the network, it is possible to obtain more flexibility to define and to (re)configure the communication links. The *Software Define Network* (SDN) approach provides flexibility to make changes during operation (for instance to define priorities). The *Network Function Virtualization* (NFV) provides independence from the hardware platform and makes it possible to easily deploy EPC functions anywhere and at any time. Both technologies can be applied to share resources in the RAN. These four technologies (FOG, MEC, SDN and NFV) can contribute and can be combined to reduce latency.

In this paper, we propose an LTE compliant architecture to reduce latency for some scenarios combining FOG, MEC and SDN. Our aim is to modify the behavior of the General Packet Radio Service (GPRS) Tunneling Protocol (GTP) tunnels to analyze the traffic in order to decide on further optimizations, for example to implement a kind of device-to-device communication for some services. This instrumentation of the GTP tunnel is compliant with the rest of the network, so we can still use Commercial-Off-The-Shelf (COTS) components in the rest of the system. Some extensions of the architecture are envisioned to take advantage of NFV in the EPC functions. Managing GTP tunnels is done with a new component between the eNodeB and the mobile network interface with the Internet.

This component, called the Fog Gateway (FGW), implements a subnetwork in the aggregation point of several eNodeBs, which includes the FGW, the User Equipment (UE) served by the FGW and local servers to optimize some services. The FGW works by inspecting and forwarding GTP packets. The control plane will work as usual and completely independently from this component which will infer the required parameters by analyzing the data plane messages. The data plane processing is based on the Tunnel Endpoint Identifier (TEID) of the GTP packets. This is done to route the IP packets to the equipment of the fog without having to traverse the entire core network. In order to speed up the process of packet inspection in the FWG, we propose the use of SDN support to filter the packets to be inspected. The solution does not imply the modification of any component in the network and it introduces a low overhead. The proposal have been validated in a real network environment at the mobile network testbed [5] of University of Málaga, where an EPC, several eNodeBs and the rest of the elements are deployed.

There are some existing papers which deal with reduction of latency in mobile networks using related approaches. The proposal in [6] also focuses on GTP tunnels to improve the service to some UEs. The authors propose a dynamic GTP termination mechanism that combines cloud based GTP with a fast GTP tunnel implemented with dedicated hardware. Depending on the requirements of the user (or other policies) the system switches from a cloud-based GTP tunnel to the fast GTP tunnel. Compared with [6], our proposal is not limited to providing one fast tunnel, and we can offer the low latency

978-1-5090-3651-6/16 $31.00 © 2016 IEEE
DOI 10.1109/FAS-W.2016.59

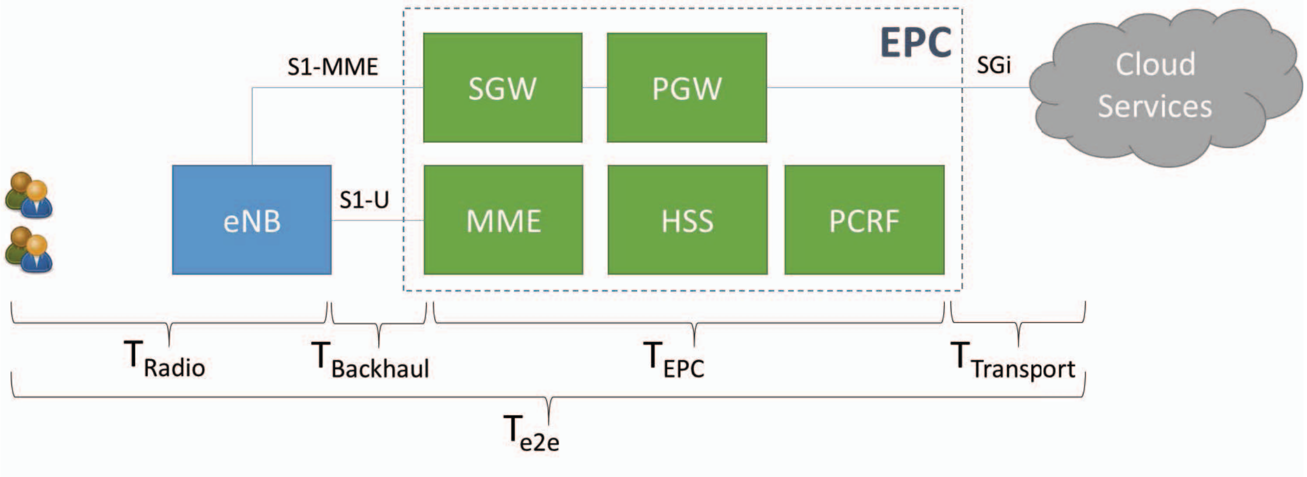


Fig. 1. LTE Basic Architecture

service to many UEs simultaneously. In addition, we do not need dedicated hardware, because the FWG and the SDN support can be implemented in the general purpose hardware available at the edge of the network. In [7] several femtocell architectures are analyzed in a qualitative fashion. The main difference from this paper is the use of components to communicate both with the control and data planes in the LTE femto architecture.

There are other proposals dealing with different aspects of latency. The proposal in [8] concentrates more on the service side. The authors analyze the business case as well as some technology and service designs. In [9] the authors present an analysis of the different architectures to support mobile edge in future 5G networks as well as some insights into the data caching and overall system performance. In [4] experimental results on NFV deployments of EPCs are given. In this approach the scalability and the setup time of the solution are analyzed thoroughly.

The paper is organized as follows. Section II presents some background on LTE networks to characterize the problem of latency. Section IV describes the architecture of the FWG subnetwork and the behavior of the new components for smart routing of user data. Section V presents the experimental platform for measuring latencies in realistic networks and validating the proposal. Conclusions and future work are discussed in Section VI.

II. Background and Problem Statement

Figure 1 depicts the basic architecture of an LTE network, where we identify several segments with different effects on latency, namely:

- T_{e2e}, is the full One Way Delay (OWD) delay, can be approximated as half the Round Trip Time (RTT).
- $T_{Backhaul}$, is the delay taken by the connection between the radio and the EPC network. This connection is normally very heterogeneous and can be based in microwaves, optical, copper solutions or combinations of them depending on the location of the base station.
- T_{EPC}, is the time taken by the components of the EPC. Most of the times the elements are connected in the same data center so the delays between them are very low.
- $T_{Transport}$, is the delay between the operator network and the cloud services.

The eNB is the LTE base station and is connected to the Evolved Packet Core (EPC) network via two interfaces, a control interface (namely S1-MME), that transport the signaling messages, and a data interface (S1-U) which carries the user data. The following subsections provide an overview of each of the components of the EPC architecture, highlighting the most relevant information as well as the latencies introduced by these elements.

A. Mobility Management Entity (MME)

The MME is in charge of the session and mobility procedures. The MME exchanges control messages with the eNB to register the users in the network and also to track the base stations where they are connected in case a handover procedure is required. In LTE the registration procedure (named attach procedure) is the most costly in terms of latency. A typical attach can take more than 1 sec. During the procedure a default Evolved Radio Bearer (eRAB) is negotiated. This radio bearer contains the transport endpoints (the designated Serving Gateway and the eNB data plane address), some identifiers and the QoS (Quality of Service) indicators. Once the mobile has been attached, an additional delay is produced when the user moves from idle (a state of low energy consumption) to connected state. This transition delay can take between 20ms and 50ms and has to be under 100ms [10].

B. Serving Gateway (SGW) and Packet Gateway (PGW)

These two components can be deployed together, the SGW maintains a control session with the MME to establish the appropriate bearers when necessary and it also provides the

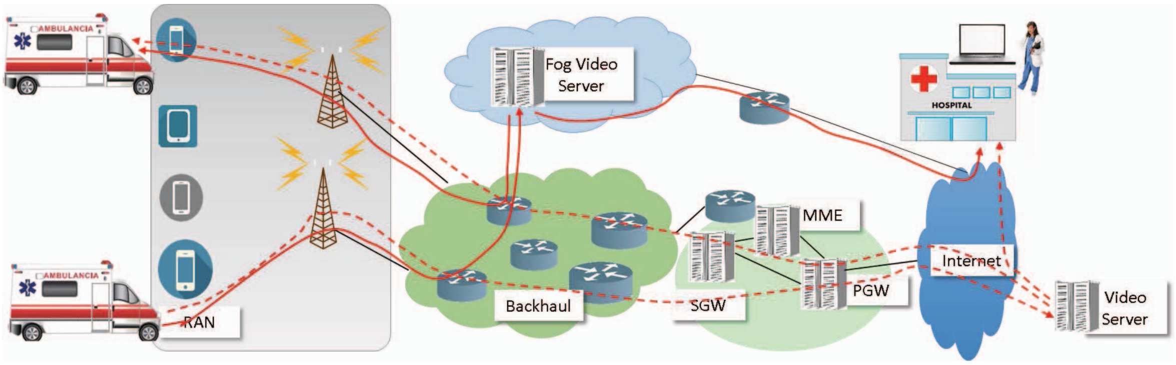
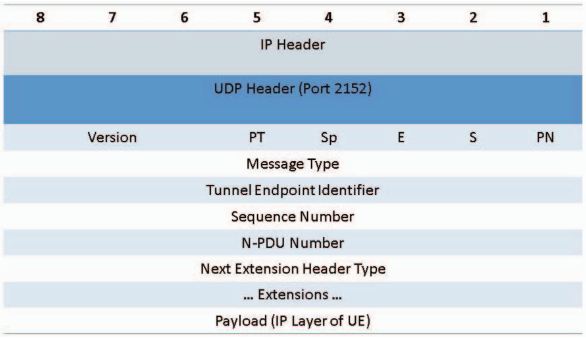


Fig. 2. Video for Emergency Communications

data plane for the eNB which is based on GTP. The PGW receives the tunnels from the SGW and routes the IP traffic of the user via the SGi interface. The GTP header (represented in figure 3 length is between 8 and 12 bytes. Its most relevant fields are the TEID and, in the case of the data messages, the payload which transports the IP layer of the UE. During the establishment in the control plane the MME and the eNB will negotiate one TEID for the default bearer uplink traffic and another one for the downlink traffic. These TEIDs together with the transport endpoints (SGW and eNB IPs) unequivocally identifies the traffic of a user in a determined transport bearer.



8	7	6	5	4	3	2	1
IP Header							
UDP Header (Port 2152)							
	Version		PT	Sp	E	S	PN
Message Type							
Tunnel Endpoint Identifier							
Sequence Number							
N-PDU Number							
Next Extension Header Type							
... Extensions ...							
Payload (IP Layer of UE)							

Fig. 3. GTP Header

C. Home Subscriber Server (HSS) and Policy and Charging Rules Function (PCRF)

The HSS is responsible for storing the user database as well as the user SIM card keys, service profiles and authorized services. It communicates with the MME via the S6a and monitors the number of times the MME accesses to the keys (this is relevant if an intermediate MME wants to be deployed). The PCRF can access the UE database and is in charge of generating rules to enforce QoS for the appropriate services. For instance the IP Multimedia Subsystem (IMS) can trigger the demand of one or more dedicated bearer for voice and/or video services.

D. Transport and End to End Latencies

The EPC components run over dedicated hardware architectures (such as ACTA) and in general terms are very expensive pieces of equipment. This a key factor that directly affects the latency. Mobile Network Operators (MNO) will normally limit the investment to a single EPC deployment (with hardware resiliency) that will be located in one of the data centers of the operator. Depending on the size of the territory to cover as well as on the transport technology used in the backhaul, the delay between the the eNB and the EPC can vary but tends to be high (in terms of tens of milliseconds). This situation is expected to worsen with 5G deployments principally due to the densification of the network that will force operators to use more, and more heterogeneous, backhaul systems. A proposal for optimizing heterogeneous backhauls and some estimated latency figures are provided in [11].

Once the UE traffic has been extracted in the PGW it should be transported from the operator SGi reference point to an Internet Exchange Point and finally to the cloud which involves another increase in the latency. We can summarize the split of the Round Trip Time (RTT) as follows:

$$T_{RTT} \approx 2 * T_{e2e} \approx T_{Radio} + T_{Backhaul} + T_{EPC} + T_{Transport}$$

In the case of peer to peer communications between users in the same eNB, the situation is worse as packets from one user to the other will have to traverse all the core network to the reference point and back:

$$T_{p2p} \approx 2 * (T_{Radio} + T_{Backhaul} + T_{EPC})$$

The objective of the FGW is to reduce these latencies keeping only those introduced by the radio, plus the overhead introduced by the new system. According to [12], the typical end to end latency is minor than 20 ms, half of them lost in the radio side and the rest split equally between the UE and the EPC. Another analysis of the latencies on live LTE networks can be found in [13]. In this case the authors analyze the end to end latency providing an average of 40ms. The equipment employed as well as the conditions of the measurements (both from the radio channel and the load of the system) can considerably affect the results of the experiments. For this

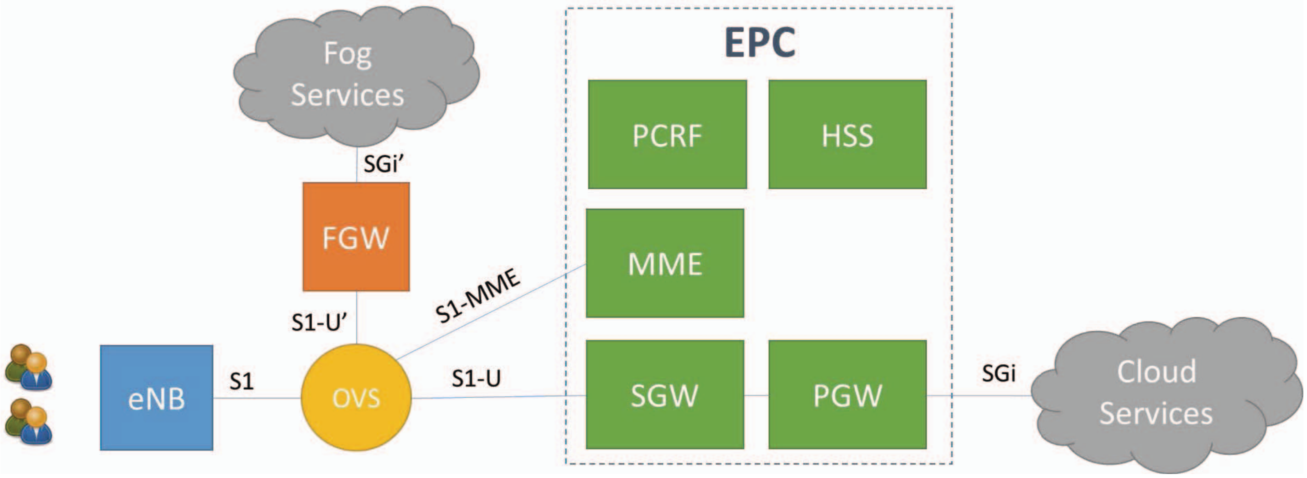


Fig. 4. LTE Proposed FOG Architecture

reason the latency breakout is analyzed in Section V so as to provide a baseline of the experimental deployment used to validate the FGW approach.

III. Motivating example

In order to motivate the use of this new component to reduce latency an example is provided. The example is based in the use case of the Q4Health [14] project, which consists on the optimization of a video platform for emergency delivery. One of the scenarios considered consists on an emergency situation where paramedics have been sent and a field command center have been deployed to coordinate the assistance of the patients. Some paramedics are equipped with a real time video streaming camera that can send images of the patients. These images will normally be sent to the command center but also to the hospitals where more specialists can provide remote support. There are several alternatives to implement the service. We can assume a video server located in the Internet domain. It can establish peer to peer connections, or the data can be sent from the video server. We can assume the later for simplicity. In that case the video from the paramedic will traverse all the mobile network, until reaching the video server via Internet and then the video server will send the video to the hospital and to the command center (which will required traversing again the core network). This is depicted in figure 2 with the dotted arrows that represents the flow of communication in the standard case.

The idea of the proposed solution is that the video server can be located close to the radio access reducing the latency to the server. The latency gains are very high in this scenario as the $T_{Backhaul}$, T_{EPC} and $T_{Transport}$ are reduced for all the peers connected to the radio access while maintaining a similar latency figure for the hospital, as depicted in figure 2 where the solid arrows represent the flows of communication in the fog scenario. Furthermore the FGW could be use to implement an on demand copy packet service that could send copy of the traffic to peers specified by an external party. This copy service could be employed to implement the before mentioned group communication but also to implement video registers, that could be useful in mission critical communications. There are more applications that could get benefit from the use of the FGW, for instance network sensors will be clearly improved, vehicle to vehicle communications and cached distribution services.

IV. Proposed Architecture

To enable low latency services on the system, we propose the introduction of an intermediate component (the FGW) between the eNB data plane and the SGW, as depicted in Figure 4. This proposed architecture also introduces an instance of Open vSwitch (OVS) to connect all the components. OVS can be used to implement rules to simplify the forwarding of the packets to the appropriate component. As a first approach we can ignore the OVS and assume that the FGW offers two S1-U' interfaces, one for the base stations and another one for the SGW. Several functions can be identified in the FGW:

- Build a user database.
- Packet routing.

The FGW offers similar interfaces to the ones provided by the combination of the SGW and PGW but without the control interface towards the MME (named S11). The S1-U' interface accepts the GTP packets of one or more eNBs and the SGi' is similar to the reference point of the LTE network, from this point IP packets can be forwarded to third parties services deployed in the operator fog subnetwork.

A. User database

In order to correctly route the fog downlink packets as well as the user to user fog communications, a database of the FGW detected users has to be built. Upon reception of a GTP package the FGW has to analyze the TEID, to determine whether or not it is from a known user. If the user is known it will be forwarded to the packet routing function and if not, the source IP addresses (both of the GTP packet and the encapsulated UE IP packet) as well as the TEID have to be stored.

B. Packet Routing

The FOG Gateway will be responsible for analyzing the GTP headers, looking for the destination of the GTP inner IP header. A basic Message Sequence Chart (MSC) with the expected behavior is shown in Figure 5. When the destination IP belongs to the fog subnetwork, the FGW will remove the GTP header and send it to the appropriate peer. In the case of packets with a destination IP outside the fog subnetwork the behavior will be the standard (the FGW will forward the GTP packet to the corresponding SGW). There are three basic cases when analyzing the GTP packet:

- Destination IP of the inner GTP IP header belongs to a server in the fog subnetwork. In this case the FGW will remove the GTP header and inject the packet towards the server. The responses from the server have to be taken, analyzed and encapsulated in the appropriate TEID.
- Destination IP of the inner GTP IP header belongs to a known User Equipment (UE) in the fog subnetwork. In this particular case belonging to the fog subnet means that FGW has detected that particular user's traffic. In this case the TEID of the packet should be modified in order to match that of the destination UE and the packet will be re-injected towards the destination eNB which may be different from the source eNB (this is the reason the source IP addresses are also stored in the database).
- Rest of the packets should be forwarded normally towards the SGW, that will route them to the Internet or to the destination eNB with no modifications.

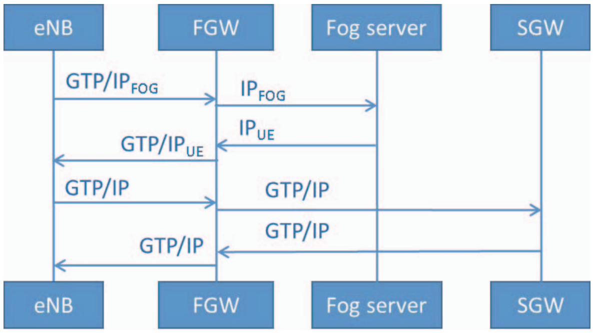


Fig. 5. MSC S1 Fog System

C. SDN optimization

The introduction of the FGW affects the latency of all the non fog services as the analysis of the GTP header and the forwarding of the packets have to be completed. In order to minimize this additional latency, SDN techniques can be applied. An instance of OVS can be modified to support the matching of packets by the inner destination IP in the GTP IP header. The packets matching the rule will be forwarded to the FGW while the rest are sent to the SGW. The database is created with the first fog packet received, which can introduce a delay with the first packets between two peers in the same node, a problem which can worsen when the communication is unicast.

D. Possible Limitations

The SGW is the termination point of the GTP tunnels, although the standard does not specify the security in the links, operators tend to implement IPsec to protect the backhaul traffic. This does not pose a problem, in the case that the FGW is used, operators will have to protect each of the segments with IPSec (from the eNB to the FGW and from the FGW to the SGW). The type of service deployed in the FGW also has to be discussed, in the case of the content services (e.g.: video distribution platforms), an analysis of the traffic might be done to decide whether or not to include the contents in cache. There are several papers covering this, such as [15][16].

The main limitation of the system will be the handover with SGW relocation, as the FGW will not see the indirect routing messages when required, and this might lead to a situation with packet duplications and/or packet losses. The base stations will be geographically close so the change of SGW will not be very frequent.

E. Expected benefits and affect on the overall system architecture

The affect of the system is very low. The GTP header analysis does not require intensive processing. It is a short header and the packets will be forwarded as they are in the GTP payload, without any modifications. In any case, as mentioned, this affect could be minimized by the use of SDN technologies in the network. The use of dedicated hardware equipment with access to the GTP header and inner IP header is expected to be available soon.

The main benefit is the decrease of the latency in the system, as it can potentially reduce the backhaul transport times and remove the transport to the server time. If this technique is combined with a dedicated bearer (to trigger a semi-persistent scheduling) the gain is even higher as the time taken to schedule and grant resources from the base station will be removed.

The following section provides an experimental validation of the approach as well as some reference measurements on the system employed to perform this validation.

V. Experimental validation

The experimental work has been developed using the PerformNetworks[1] testbed. The elements employed are:

- T2010 conformance testing equipment with S1 extensions (to enable connection to standard EPCs). This equipment is used to supports design verification of UEs. For this experiments a modified version that support standard S1 communications has been employed. With this equipment we can emulate a base station and also the effect of the

[1] http://performnetworks.morse.uma.es

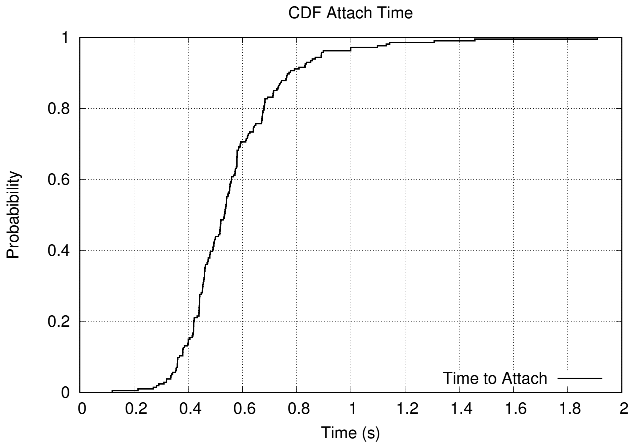


Fig. 6. Attach Time in seconds (CDF)

channel, as it is able to emulate different profiles of fading and noise.

- Commercial small cells, used as a reference COTS equipment.
- EPC emulator, which is a software that can emulate one or more instances of a basic core network still providing carrier grade performance for a small group of users (less than 300). Additionally the system supports the generation of impairments in any element interface. This functionality has been employed to emulate the effect of the backhaul and transport networks.
- Several tools that have been specifically designed to analyze the latency in the different components of the system.

More detail on the different components of the testing platform are provided in [5]. To provide more stable clocks, all the components of the system have been synchronized using a stratum 2 Precision Time Protocol (PTP) server based on Global Positioning System (GPS) system.

A. Control Plane Baseline

Although the optimization of the control plane is not covered by the FGW system it is important to estimate some values on how long it takes, as this is relevant for many of the potential use cases of the platform (such as IoT or Vehicle to Vehicle communications). A tool for analyzing S1 traces have been implemented. The tool estimates the attach establishment time by analyzing the Non Access Stratum (NAS) signaling messages between the UE and the eNB. The estimation is calculated as the difference between the NAS Attach Request and the NAS Attach Complete messages. This estimation does not consider random access procedure time, which is highly dependent on the number of users [17], so up to an additional 100ms could be taken into account.

The results provided are based on the analysis of traces of the testbed which covers attach procedure in many different scenarios (live deployments, emulated channel conditions, commercial equipment, open source eNB implementations, etc.). The main limitation is the number of users in the cell which in all the cases has been low (no more than 10 simultaneous users and frequently a single UE) and the backhaul transport time (most of the deployments have no backhaul network). Figure 6 depicts the Cumulative Distribution Function of the Attach Setup time, more than 800 samples of the attach procedure have been analyzed. The median setup time of the attach procedure is 534.498 ms with a median absolute deviation of 94.367 ms.

TABLE I
RTT SPLIT TIME COMPARISON BETWEEN SMALL CELL AND T2010

	COTS Small Cells		T2010 Equipment	
Time	Median RTT	MAD RTT	Median RTT	MAD RTT
T_{e2e}	28.775 ms	4.887 ms	11.830 ms	0.253 ms
T_{Radio}	28.223 ms	4.882 ms	11.577 ms	0.247 ms
T_{EPC}	0.227 ms	2 us	0.229 ms	3 us

TABLE II
T2010 CONFIGURATION PARAMETERS FOR THE EXPERIMENTS

Parameter	Configured Value
Frequency (Band 20)	DL 806MHz UL 847MHz
Bandwidth	10 MHz
Power	-61 dBm/15KHz
Modulation	22-64QAM (Both UL and DL)
Max. HARQ Retransmissions	7

B. Data Plane Baseline

The measurements in this section have been taken based on the analysis of ping traces. A tool has been developed to analyze the messages in the UE, SGi interface and the S1-U. And has been used to compare the measurements on the network, as well as to have estimations of T_{Radio}, T_{EPC} and the RTT (T_{RTT}).

Two campaigns of measurements for the data plane have been performed. In the first campaign we measure the latency breakout in the different components of the system in an ideal environment both for a small cell and the T2010 emulator. Table I provides an overview of the split of the measurements in the system. In our system, most of the time, the end to end delay is lost in the radio link, the figures of the EPC employed are very low due to the fact that the number of users is very low (fewer than ten) compared to a commercial one (can scale to millions). The T2010 employs a fixed scheduling policy as it only supports a single UE, which explains the difference in the end to end latency.

The next campaign of measurements consists in the emulation of a realistic network using the T2010 system. To do so we have included impairments on the EPC side both for the transport and backhaul times and on the radio side by the introduction of impairments and noise in the radio links. The configuration of the T2010 is summarized in Table II. The mean delay of the backhaul is 15 ms and 30 ms for the transport. Additionally the radio conditions have also been simulated to include good conditions (no fading, nor noise), medium conditions (vehicular fading and noise power

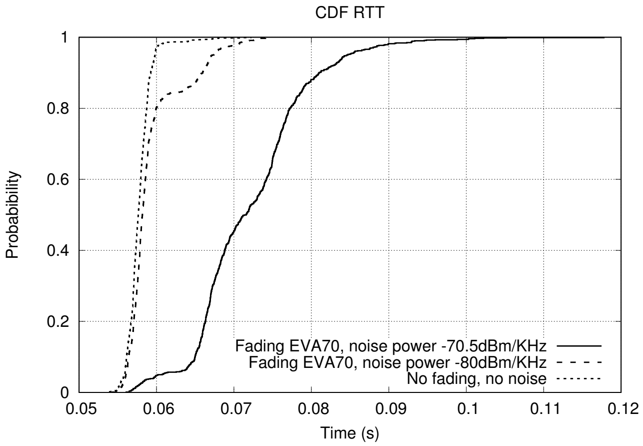


Fig. 7. T2010 RTT (s) baseline with different channel conditions (CDF)

TABLE III
SUMMARY OF THE RESULTS OF THE BASELINE

	Conditions		
	Ideal	Medium	Bad
RTT_{e2e}	57.642 ms	58.179 ms	71.547 ms
$BLER_{MAC}$	0 %	10 %	50 %

-80dBm/15KHz) and bad conditions (vehicular fading and noise power -70.5dBm/KHz). Figure 7 depicts the Cumulative Distribution Function (CDF) for the three different channel conditions considered in the experiment. We note that when the radio conditions get worst the latency increases, the main reason is the retransmissions that are triggered at MAC level. We took some measurements of the BLER (at link level), for the case of no noise it was 0%, 10% in the case of the -80dBm/15KHz and almost 50% in the case of the -70.5dBm/15KHz. In all the cases the end to end BLER was 0% due to the LTE Hybrid Automatic Repeat Request (HARQ) retransmission procedures. A summary of the results is provided in Table III.

C. FGW Results

In order to validate the approach being used we compared the baseline results with the results of collocating the functionality of the SGW to the eNB. This scenario provide a performance similar to the one expected by the deployment of an FGW without some of the limitations (which anyway do not apply in the scenario being used). Figure 8 presents the comparison between the delays with the fog and the regular traffic. The measurement conditions are the same employed for the baseline except that the scenarios have been mixed (we have tested in the same experiment with changes between bad, medium and good channel conditions). Table IV provides a summary of the results obtained, the latency is reduced by up to 78% percent. This gain depends highly on the latencies considered in the system which will normally depend on the geographical location (which will determine the distance to the EPC) and also the type of backhaul that can be employed.

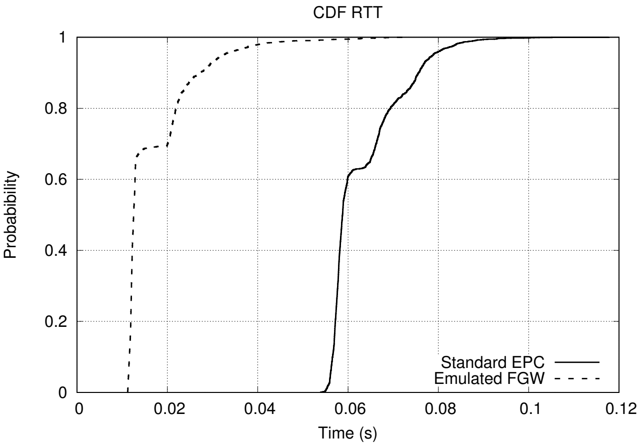


Fig. 8. RTT (s) comparison between emulated FGW and baseline (CDF)

TABLE IV
SUMMARY OF THE COMPARISON BETWEEN THE FGW AND THE BASELINE

	Emulated FGW	Baseline
Median RTT	12.5 ms	58.8 ms
MAD RTT	6.2 ms	6.7 ms

Additionally the trade off of the system has not been measured, the affect of the FGW system on the non fog traffic has had to be estimated but, taking into account the time split between the different components, the effect will be negligible.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented an optimization method to reduce latency for communications among users in the same area of the mobile network. The fog gateway has been evaluated in a realistic environment and we have obtained results that confirm the applicability in a real network without major modifications. The effect on the latency can be very positive depending on the type of deployment. More analysis is still to be done, especially in terms of scalability to determine the affect of the component when the number of users is very high and to discover the optimal number of stations to be used in a single FGW.

The use in content based services (such as video distribution, web pages, etc.) has yet to be studied, the caching of contents might limit the gains in latency but on the other hand the use in user to user communication is very suitable. For instance the FGW could be deployed in the base stations close to highways to provide low latency vehicle to vehicle communications.

We are considering several lines of research for future work. The first task will be to generate more complex scenarios such as live network deployments and interactions with an operator EPC. This new measurement campaign will analyze the effect of the gateway to support user to user real time video communications. We are also exploring a similar approach for the control plane, in order to reduce the time of the network procedures (like attach, handover, etc.). As described in section

IV the effect of SDN can be very positive and can reduce the tradeoff introduced by the FGW in the system, so the use of modified versions of OVS will also be explored.

Acknowledgment

The work described in this paper has received funding from the European Horizon 2020 Programme under grant agreement 688624 (Q4Health project).

References

[1] N. Alliance, “NGMN 5G White Paper,” Next Generation Mobile Networks, Tech. Rep., Feb. 2015.

[2] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, “Fog computing and its role in the internet of things,” in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC ’12. ACM, 2012, pp. 13–16. [Online]. Available: http://doi.acm.org/10.1145/2342509.2342513

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, “ETSI White Paper No. 11. Mobile Edge Computing A key technology towards 5G -White paper,” ETSI, Tech. Rep., Sep. 2015.

[4] E. Cau, M. Corici, P. Bellavista, L. Foschini, G. Carella, A. Edmonds, and T. M. Bohnert, “Efficient exploitation of mobile edge computing for virtualized 5g in epc architectures,” in *2016 4th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, March 2016, pp. 100–109.

[5] A. Díaz Zayas, C. García Pérez, and P. Merino Gomez, “PerformLTE: A testbed for LTE testing in the future internet,” in *Wired/Wireless Internet Communications: 13th International Conference, WWIC 2015, Malaga, Spain, May 25-27, 2015, Revised Selected Papers*, vol. 9071. Springer, 2015, p. 46.

[6] J. Heinonen, T. Partti, M. Kallio, K. Lappalainen, H. Flinck, and J. Hillo, “Dynamic tunnel switching for sdn-based cellular core networks,” in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, &#38; Challenges*, ser. AllThingsCellular ’14. ACM, 2014, pp. 27–32. [Online]. Available: http://doi.acm.org/10.1145/2627585.2627587

[7] F. Lobillo, Z. Becvar, M. A. Puente, P. Mach, F. L. Presti, F. Gambetti, M. Goldhamer, J. Vidal, A. K. Widiawan, and E. Calvanesse, “An architecture for mobile computation offloading on cloud-enabled lte small cells,” in *Wireless Communications and Networking Conference Workshops (WCNCW), 2014 IEEE*, April 2014, pp. 1–6.

[8] O. Mäkinen, “Streaming at the edge: Local service concepts utilizing mobile edge computing,” in *Next Generation Mobile Applications, Services and Technologies, 2015 9th International Conference on*, Sept 2015, pp. 1–6.

[9] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, “Edge cloud and underlay networks: Empowering 5g cell-less wireless architecture,” in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, May 2014, pp. 1–6.

[10] N. Maskey, S. Horsmanheimo, and L. Tuomimäki, “Latency analysis of lte network for m2m applications,” in *Telecommunications (ConTEL), 2015 13th International Conference on*, July 2015, pp. 1–7.

[11] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, “Fundamentals of heterogeneous backhaul design, analysis and optimization,” *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 876–889, Feb 2016.

[12] N. S. Networks, “The impact of latency on application performance,” Nokia Siemens Networks, Tech. Rep., 2009.

[13] M. Laner, P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato, and M. Rupp, “A comparison between one-way delays in operating hspa and lte networks,” in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on*, May 2012, pp. 286–292.

[14] C. A. Garcia-Perez, A. Rios, P. Merino, K. Katsalis, N. Nikaein, R. Figueiredo, D. Morris, and T. O’Callaghan, “Q4health: Quality of service and prioritisation for emergency services in the LTE RAN stack,” in *Networks and Communications (EuCNC), 2016 European Conference on*, June 2016.

[15] J. Oueis, E. C. Strinati, and S. Barbarossa, “The fog balancing: Load distribution for small cell cloud computing,” in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–6.

[16] A. S. G. T. Braun and E. Monteiro, “Enhanced caching strategies at the edge of lte mobile networks,” in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, May 2016, pp. 341–349.

[17] K. Zhou and N. Nikaein, “Low latency random access with TTI bundling in LTE/LTE-A,” in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 2257–2263.