Definition and Evaluation of Latency in 5G: A Framework Approach

K.X. Du
OnApp
UK

G. Carrozzo
Nextworks
Italy

M.S. Siddiqui
Fundació i2CAT
Spain

O. Carrasco
Casa Comms.
Spain

B. Sayadi
Nokia Bell Labs
France

F. Lazarakis, A. Kourtis
NCSR Demokritos
Greece

J. Sterle
Internet Institute
Slovenia

R. Bruschi
CNIT
Italy

***Abstract*— Latency as a key performance indicator (KPI) of 5G communication has attracted a significant amount of efforts from both industry and academia to provide the enhanced technology solution to meet the requirement of low latency from the diverse 5G use cases. With the evolution of technologies and eco-system in 5G, it becomes not only more important but also more challenging to manage a large number of elements which have been connected and orchestrated to provide an end-to-end communication service across different segments of the network. This paper introduces a framework approach to define and evaluate latency in 5G based on the proposed 5G systems from four research projects. The use case and architecture are well studied to identify and analyse the contribution of latency. The proposed reference framework of latency is validated by mapping it to the testbed of these projects.**

***Index Terms*— 5G, latency, Key Performance Indicators, evaluation**

I. INTRODUCTION

This paper is a joint work among four 5G-PPP Phase 2 projects, namely, NGPaaS [1], 5GCity [2], 5G ESSENSE [3] and MATILDA [4]. The motivation of this join work is driven by the significant diversity in both use cases and technologies of 5G. In order to meet the requirements from different vertical use cases, the 5G deployment shows high heterogeneity in virtualisation technologies, platform and workload deployment etc., which range from the physical infrastructure to the upper-layer services tailored to the vertical services. Thus, the collaboration among multiple projects could be precious to capture and cope with different views from these projects. The goal of this joint work is to provide a methodology to define and evaluate the latency as a KPI of 5G communication under different scenarios and technical solutions.

In order to select a suitable technology solution to meet the latency requirement of use cases, it is necessary to have a reference framework to facilitate the evaluation and comparison between the different solutions. The definition of latency is an critical milestone to produce the reference framework. This paper presents a reference framework of latency as an approach to not only define the latency but also evaluate the latency, which is well studied based on the testbed of these projects.

The remainder of this paper is organised as follows. Section II reviews the latency-critical use cases addressed by the solutions proposed by these projects, including their eco-system and the challenges. The components and testbed for these use cases are presented in Section III. The definitions and measurement methodologies of latency are provided in Section IV, followed by the solutions for reducing latency and the corresponding evaluation approach in Section V. Finally Section VI concludes this paper and provides the future work.

II. Latency-critical Use Cases

The Ultra Reliable Low Latency Communications (URLLC) is one important emerging area in 5G communication to support latency-critical services [5]. In this section, the latency-critical use cases targeted by different projects are presented under their novel architecture or eco-system.

A. *MCPTT Service Provider backed by NGPaaS Operator*

The MCPTT (Mission Critical Push to Talk) use case is inspired by the mission critical use case defined by 3GPP and targeted to provide group call services by an MCPTT service provider to its users (i.e. police forces, public safety organisations, etc.). The NGPaaS project works on one scenario of MCPTT use case, which is in a very large factory that has been set on fire, firefighters arrived at the site to extinguish the fire. Firemen are divided into several groups to fight against the fire and evacuate people with the support of intensive voice communication between each other to synchronise and distribute orders in real-time. The MCPTT use case involves three actors: the NGPaaS operator, the MCPTT service provider and the MCPTT user. The NGPaaS operator provides a PaaS (Platform as a Service) consisted of the infrastructure and virtualised network functions (VNFs) to set up a connectivity for communication while the MCPTT service provider can request and consume this connectivity on demand to run the MCPTT Apps, which are presented as the end service to be consumed by the MCPTT user, e.g. firemen in the example above.

The architecture of the MCPTT use case based on this firefighting example is depicted in Fig. 1. The fire truck is

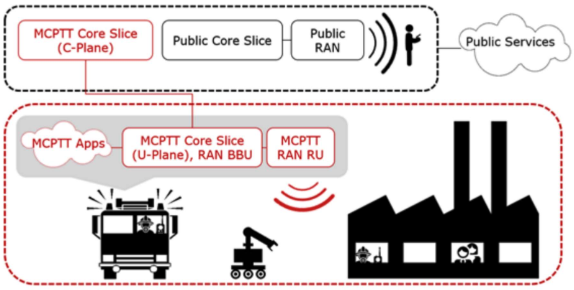


Fig. 1. MCPTT Use Case in NGPaaS: firefighting communication

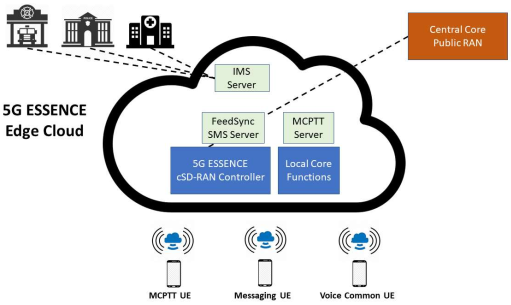


Fig. 2. Edge Cloud of 5G ESSENCE Public Safety Use Case

equipped with the infrastructure as an edge cloud with the capability to host User Plane (UP) components of Core Network (CN), components of the radio access network (RAN) and MCPTT Apps. The Control Plane (CP) components of the Core Network are hosted in a central cloud. Except for the MCPTT Apps, all the other components are provided and maintained by NGPaaS operator as a PaaS.

B. Public Safety Use Case of 5G ESSENCE

Following the most relevant trends in terms of a use case definition for URLLC, together with Multi-Tenancy management, 5G ESSENCE project has identified the Public Safety use case as one of the most representative examples for URLLC application. The process of providing the MCPTT service in 5G ESSENCE can be summarized as follows. 5G ESSENCE infrastructure operator provides the required network slices to different tenants (Mission Critical Organizations) with its respective SLAs. Allocation of QoS (Quality of Service) in each slice is guaranteed by the cSD-RAN (centralized Software Defined - Radio Access Network) controller in accordance with the cloud resources allocated in the 5G ESSENCE Edge Cloud, where a set of Cloud Enabled Small Cells (CESCs) provides RAN resources with close-to-zero delay, maintaining the network services even if the backhaul is damaged, enforcing the priority access of first-responders creating the end-to-end slices that isolate those responders from other Mission Critical organizations. In case of damaged infrastructure, 5G ESSENCE Edge Cloud infrastructure maintains the service operation terminating both the control plane and the data plane in the edge by deploying local core functions VNFs. As depicted in Fig. 2, the 5G ESSENCE Edge Cloud can manage low latency services deployed in the edge network, being able to route to the different Mission Critical Organizations the messaging, data and voice communication that allows providing to the different public safety teams an efficient coordination besides the specific local services needed for a comprehensive solution to serve both first responders and public safety teams.

C. Media Use Cases for Smart Cities

The media industry use cases are particularly relevant to the Smart City environment, due to the increasing diffusion among citizens of UHD (Ultra High Definition) streaming and immersive media services. In the UHD video distribution and immersive services use case (illustrated in Fig. 3), the service typology for media distribution and immersive experience is on-demand and the media consumption takes place on the move, in mobility across the city areas covered by the 5GCity network. Various types of devices (e.g. smartphones, tablets and virtual reality devices) are used and various sections of the virtualised network infrastructure are dynamically configured to provision the service, which involves the media servers in 5GCity metro nodes/datacenters, edge computing nodes for local transcoding and video caching, and far edge computing for RAN virtualisation. When the immersive aspects come into play, additional contents need to be automatically retrieved from the media server/libraries at metro datacenter or in edge caches, in the form of 2D video, panoramic video and 3D models to augment the reality where the user is immersed. In this case, low end-to-end latency can allow high responsiveness of the immersive application functions.

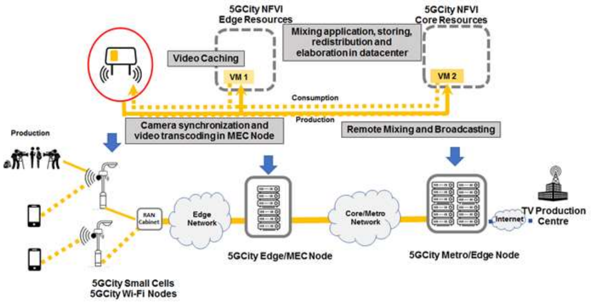


Fig. 3. UHD streaming and immersive services in 5GCity

D. Mission Critical Data in Disaster Relief in MATILDA

Catastrophic events such as earthquakes attracted the community's attention on the need for powerful and resilient emergency communication networks. Meanwhile, the 3GPP extensions for mission critical data (MC-Data) services and applications are maturing into standards to support future Public Protection and Disaster Relief (PPDR) communications. The Mission Critical Data in Disaster Relief (MC-

DRR) scenario in the MATILDA project makes use of this capability to deliver a suite of low-latency services and applications on top of a 5G infrastructure. The services are designed for emergency response teams both in day-to-day operations and during extreme situations requiring large on-site interventions and support real-time intervention monitoring as well as a series of mobility and location tracking capabilities that can be used during emergency operations of various scales.

To support MC-DDR use case, the following actors and stakeholder are part of the MATILDA based 5G emergency ecosystem. The BB-PPDR (BroadBand-PPDR) network operator provides and operates MATILDA services and cloud infrastructure. Each emergency response organization (ERO) has a dedicated BB-PPDR service provider functioning as MATILDA service provider. Emergency response teams and end users (policemen, firefighters, EMT members) are in the role of MATILDA service consumers.

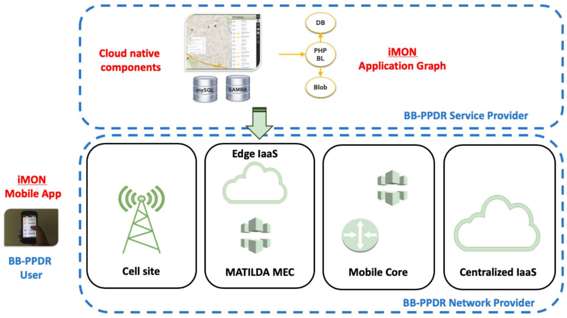


Fig. 4. Deployment of MC-DDR application for real-time intervention monitoring (iMON) at Edge Cloud

The targeted KPIs for MC-DDR use case include low latency capabilities of 5G user and control plane, along with deployment automation, high availability, resilience and system flexibility. To extend the system for the most extreme MC-DDR applications (e.g. remote drone control) sub 1 ms latency of user plane is required.

III. Testbed and Components

In order to demonstrate the solutions which could be beneficial to the aforementioned use cases, each project has the testbed and prototype system, which are useful for us to better understand the segments of latency by looking into the components and its connectivity.

A. MCPTT in NGPaaS

As illustrated in Fig. 5, the MCPTT system is deployed in two different locations as central and edge cloud, where two Kubernetes clusters are deployed respectively. All the Kubernetes nodes are deployed in VMs managed by OpenStack except the node used to run the RAN components, which can benefit from the native bare metal performance to achieve low latency. In addition, a USRP board is connected to the bare-metal node in edge cloud to act as the antenna and provide the radio interface for connection to the UE (User Equipment). In the MCPTT use case, the PaaS needs to support very low latency, which will require raising the bar in terms of performance, especially at the RAN level. In addition, the deployment time is also critical in this MCPTT use case. These requirements are the motivation to have modularity, build-to-order design principle and various Telco-grade enhancements in Kubernetes [6].

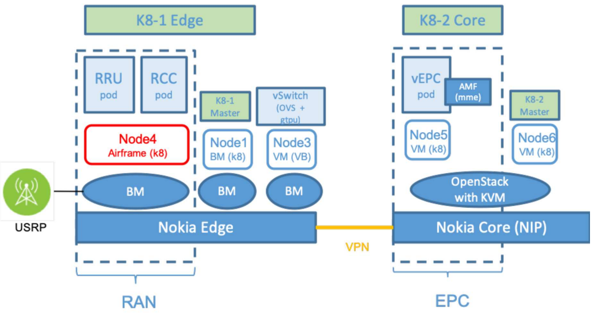


Fig. 5. MCPTT testbed deployment in NGPaaS project

B. MCPTT in 5G ESSENCE

The Fig. 6 shows the different components involved in the 5G ESSENCE solution for Public Safety services, which include an MCPTT core VNF enabling the communication using the MCPTT app, the IP Multimedia Subsystem (IMS) core implementing the local voice communications linking the MCPTT service with the Mission Critical Organisations, the localisation and messaging service implemented using a VNF deployed in the Edge Cloud called FeedSync, which is a subscriber-based content distribution tool based on an innovative modular solution that operates on top of the 5G ESSENCE, and Local Core functions that allows to implement a decentralized Core function for supporting end-to-end services terminating both the data plane and control plane in the edge network.

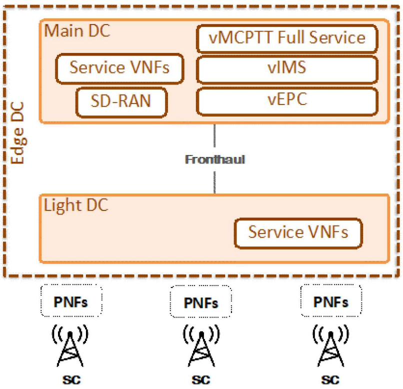


Fig. 6. Components in 5G ESSENCE solution for Public Safety services

C. UHD Streaming and Immersive Services in 5GCity

The environment for execution and measurement of the KPIs in the UHD streaming and immersive services by 5GCity is presented in Fig. 7. The infrastructure is composed by a 5GCity Metro Node hosted in the datacenter, a 5GCity MEC node hosted in a city cabinet, and two clusters of radio resources implemented as Small Cells. All components are interconnected by fibre network owned by the local municipality of the trail city. One use case uses a slice of the infrastructure which is formed by VNFs in the compute nodes, network resources and a dedicated channel in the Small Cells.

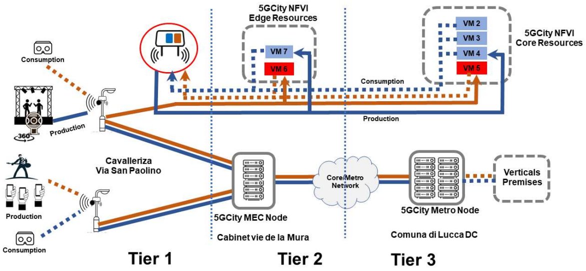


Fig. 7. Planned pilot for the KPI validation of UHD streaming and immersive services in the 5GCity infrastructure in Lucca

D. MC-DDR in MATILDA

The MC-DDR use case has been deployed in one of the three demonstration facilities of the MATILDA project. The testbed consists of software-driven hardware components (i.e. servers, SDN switches and Software-defined Radio (SDR) devices), which are centrally managed through an ad-hoc Metal-as-a-Service (MaaS) controller. Within the MATILDA framework, the MaaS controller automatically deploys a number of OpenStack platforms to serve as 5G-aware Telco multi-IaaS environment, the SDN transport networks and different physical functions including SDR-based eNodeB. In addition, a number of instrumentation and emulation components (e.g. hardware-based Ixia traffic generators and emulators) have been available to conduct latency measurements. The clock is synchronized among all the components by means of the IEEE 1588 protocol with the hardware-assisted timestamp.

As shown in Fig. 8, the reference environment applied in the MATILDA testbed consists of two edge and one core OpenStack VIMs. The reference NFV service is a softwarized 3GPP 4.5G network with the addition of a bump-in-the-wire VNF. As suggested by the ETSI MEC working group, this VNF is used to connect edge computing facilities before the 4G Enhanced Packet Core. Through this setup, MATILDA is investigating how latency can affect the end-to-end operational behaviour of vertical applications deployed within the network VIMs by considering the performance and dynamics of both user and control plane.

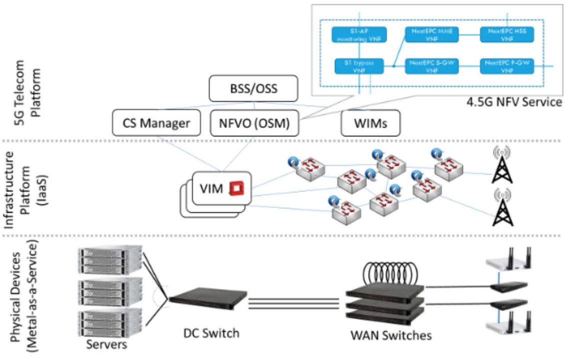


Fig. 8. Main building blocks of the MATILDA testbed at CNIT-S3ITI facilities.

IV. Definition and Measurement of Latency

In order to analyse the latency and improve the latency of the latency-critical use cases introduced in the previous section, it is necessary to figure out the components and connectivity between them. Based on the testbed and components of each use case proposed by the project, a reference framework is proposed to facilitate the definition and measurement of latency when the different solutions are provided and compared.

A. Reference Framework for Latency in 5G

From the proposed testbed and components from each project we can see that the system architecture follows the key principles of the 3GPP TS 23.501, i.e. separating the User Plane functions from the Control Plane functions, allowing independent scalability, evolution and flexible deployments, e.g. centralized location or distributed (remote) location. In this work, the focus is on user plane latency. The components including the VNFs used to build the 5G system are illustrated as a service-based architecture in Fig. 9. The VNFs involved in the user plane are UE, RAN, User Plane Function (UPF) and Data Networks (DN).

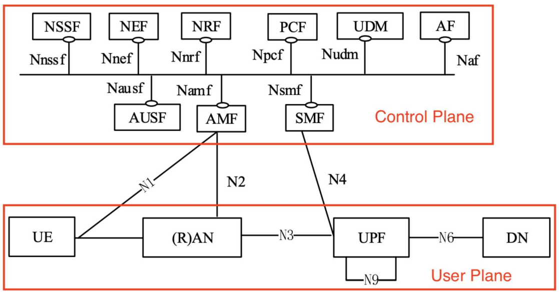


Fig. 9. Service-based Architecture with separation of Control Plane and User Plane

Based on the analysis of the components and the contribution of latency in the user plane, a reference framework from user-plane latency aspect is proposed as the Fig. 10, where the latency is considered to be composed of three types of

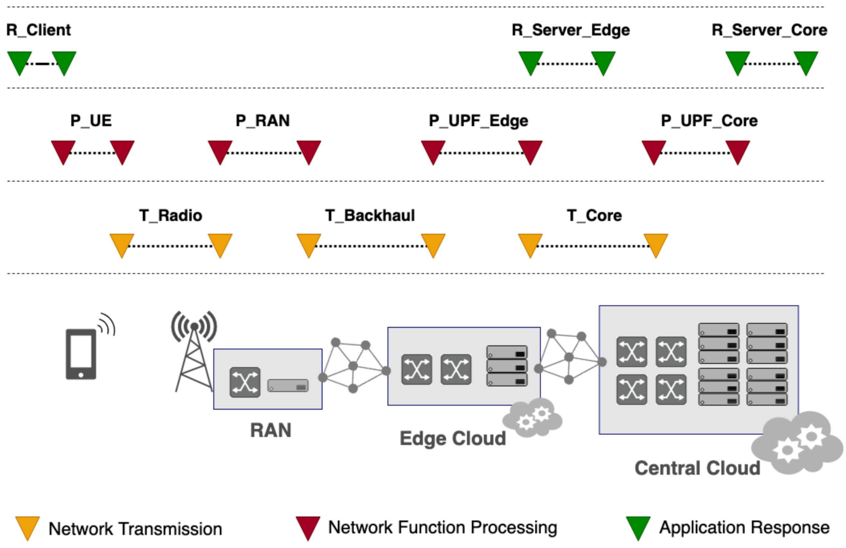
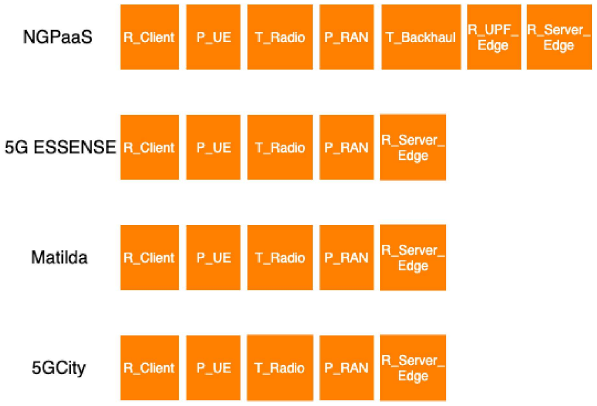


Fig. 10. Reference framework of delay contributions of an end-to-end latency

delay: network transmission time (depicted as T), network function processing time (depicted as P) and application response time (depicted as R). For example, P_{UE} denotes the processing time of a packet since it is sent from the UE's application layer until it is transmitted by the UE's physical layer, while T_{radio} denotes the network transmission time from the egress of UE to the ingress of RAN.

B. Mapping of Reference Framework

The latency addressed in each project is defined based on the reference framework to specify the scope of latency measured in each testbed, which is illustrated in Table. I and Fig. 11. Among the segments of latency contributions, some of them are enhanced using the solutions introduced in these projects.



NGPaaS R_Client P_UE T_Radio P_RAN T_Backhaul R_UPF_Edge R_Server_Edge

5G ESSENSE R_Client P_UE T_Radio P_RAN R_Server_Edge

Matilda R_Client P_UE T_Radio P_RAN R_Server_Edge

5GCity R_Client P_UE T_Radio P_RAN R_Server_Edge

Fig. 11. The latency addressed by each project in the reference framework

V. Evaluation with Reference Framework

Each project has been working on different solutions to reduce the latency from different segments as defined in the previous section. With the availability of the reference framework, it enables the evaluation and comparison with each other in a more structured approach, which helps any new solution to be positioned and recognized easily. The proposed solutions in the above projects are provided in Table II with the different challenges that could be addressed by these solutions.

TABLE II
Solutions to Reduce Latency

Challenges	**Solutions Developed in Projects**	**Segment(s) in Reference Framework**
Performance shortage of container-based NFVI for RAN Components	NUMA-aware CPU Pining for Kubernetes [7] FPGA Device Plugin for Docker and Kubernetes [8] Partial Reconfiguration for FPGA Virtualisation [8]	P_RAN
Long response time in Mission Critical services	Implementing a distributed MCPTT server that can operate on multiple Edge DCs [10].	R_Server_Edge
End-to-end delay for media services	Resource orchestration and allocation to optimize network and application latency Functional split of the application core functions and distribution across the core and edge section of the 5GCity infrastructure [9]	R_Server_Edge
End-to-end latency requirements for interactive applications	Deployment of cloud service components and 5G network elements at Edge Cloud (MEC) [10]	P_RAN R_Server_Edge

TABLE I

Definitions and Measurements of Latency

Project	Use Case	Latency Definition	Component Setup	Measurement Tool
NGPaaS	MCPTT	The time since the packets are sent from a UE to an application server located at the same place of RAN, until the application received the packets.	A client PC with an antenna device sends packets to a server where OAI software is running on top of infrastructure using Kubernetes with NUMA-aware CPU pinning feature.	PING/ICMP
5G ESSENSE	MCPTT	Latency is measured as The End-to-end MCPTT Access Time. The time between when an MCPTT user requests to speak and when this user gets a signal to start speaking, including acknowledgement from first receiving user before voice is transmitted.	An MCPTT broadcast call is performed from the UE in a periodic manner. The MCPTT setup call time is measured and stored in the 5G ESSENCE Cloud-Enabled Small Cell Manager (CESCM).	MCPTT Call Setup Time
5G City	UHD Video Streaming	Multiple intermediate measurement points: a) UE - Edge computing instance; b) Edge - core data center network; c) [end-to-end] UE - application server.	A UE and an application server running at core data center, the edge (e.g street cabinet) and far edge (e.g. lamppost) where network functions are allocated.	PING/ICMP
MATILDA	MC-DDR	Round-trip time delay on the connected UE. Measurement reference packet is sent and its response is received by the same UE device.	A UE and application components. Measurements results are stored in Prometheus platform and reported to the MATILDA OSM and VAO.	PING/ICMP

TABLE III

Evaluation of Latency

Latency	Evaluation Methodology
Latency on RAN Service running on Kubernetes	Comparing the latency in a relative approach based on different setup: No CPU pinning (RRU and RCC pods are setup with Best Effort K8S QOS class), CPU pinning on the wrong NUMA node regarding USB Airframe card (RRU and RCC pods are setup with Garanteed K8S QOS class with an annotation to select the wrong NUMA), and CPU pinning on the right NUMA node.
Data plane round-trip time latency for MC-DDR applications	Targeting the absolute number: End-to-end round-trip time latency for mission critical applications must be less than 1 ms; End-to-end round-trip time latency for interactive applications must be less than 20 ms.
E2E latency across multiple nodes from UE to application server.	Targeting the absolute number: a) UE - Edge computing instance $\leq$ 1 ms (depends on load, UE distance from small cell, propagation conditions, does not include processing time at network functions) b) edge - core data center network latency $\leq$ 2 ms (typically on fiber network) c) end-to-end client - application server one-way delay $\leq$ 10 ms (full scope defined in Fig. 10)

The latency addressed by the above solutions is measured using the setup and tools introduced in Table I, while how the measured results are evaluated is presented in Table III.

VI. Conclusion And Future Work

This paper introduces a framework approach proposed and followed by four research projects to define and evaluate the latency within the context of different latency-critical use cases. This reference framework enables the researchers and developers working on latency-relevant solutions to analyse and illustrate the latest innovation and technology achievements in a more formalised approach, which could result in the more clear and strong impact of novel technologies for reducing latency in the 5G community.

Future work will consist of consolidating individual measurement strategies, categorising the latency-reduction solutions based on specific segments of the latency contribution as per Fig. 10, progressing the evaluation and sharing the results.

ACKNOWLEDGMENT

This work is supported by four H2020 projects within the H2020 research and innovation program for Phase-2 of 5G Infrastructure Public-Private Partnership: NGPaaS (GA-761557), 5G ESSENCE (GA-761592), 5GCity (GA-761508) and MATILDA (GA-761898).

References

[1] The ngpaas project website. [Online]. Available: http://www.ngpaas.eu/

[2] The 5gcity project website. [Online]. Available: http://www.5gcity.eu/

[3] The 5g essence project website. [Online]. Available: http://www.5g-essence-h2020.eu/

[4] The matilda project website. [Online]. Available: http://www.matilda-5g.eu/

[5] I. Parvez, A. Rahmati, I. Güvenç, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5g: Ran, core network and caching solutions," *CoRR*, vol. abs/1708.02562, 2017.

[6] A. Mimidis, E. Ollora, J. Soler, S. Bessem, L. Roullet, S. Van Rossem, S. Pinneterre, M. Paolino, D. Raho, X. Du, J. Chesterfield, M. Flouris, L. Mariani, O. Riganelli, M. Mobilio, A. Ramos, I. Labrador, A. Broadbent, P. Veitch, and M. Zembra, "The next generation platform as a service cloudifying service deployments in telco-operators infrastructure," in *2018 25th International Conference on Telecommunications (ICT)*, June 2018, pp. 399–404.

[7] E. B. e. a. Michele Paolino, Marco Mobilio, "Deliverable 4.1-telco-grade paas: First results and implementation," 2018. [Online]. Available: http://ngpaas.eu/projectoutcomes/ngpaasdeliverables

[8] R. Kerherve, J. Lallet, L. Beaulieu, I. Fajjari, P. Veitch, J. Dion, B. Sayadi, and L. Roullet, "Next generation platform as a service: Toward virtualized dvb-rcs2 decoding system," *IEEE Transactions on Broadcasting*, pp. 1–9, 2019.

[9] S. e. a. Khalili, Hazmeh; Papageorgiou; Siddiqui, "5gcity project deliverable d4.1," 2017. [Online]. Available: https://zenodo.org/record/2558306.XOFpN8gzaUk

[10] R. Bruschi, F. Davoli, P. Lago, and J. F. Pajo, "A multi-clustering approach to scale distributed tenant networks for mobile edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 499–514, March 2019.