Definition and Evaluation of Latency in 5G with Heterogeneous Use Cases and Architectures

Gino Carrozzo
Nextworks
Italy

M. Shuaib Siddiqui
Fundació i2CAT
Spain

Kevin Du
OnApp
Gibraltar

Bessem Sayadi
Nokia Bell Labs
France

Oscar Carrasco
Casa Comms.
Spain

Fotis Lazarakis
NCSR Demokritos
Greece

Janez Sterle
Internet Institute
Slovenia

Roberto Bruschi
CNIT
Italy

***Abstract*— Low latency communications are a critical enabling factor of various 5G use cases. Since their initial conceptions, 5G technologies for the various elements of the network (i.e. air interface, fronthaul, backhaul, edge, core, transport) have been designed to pursue a drastic reduction of latency at the various segments of the communication service and eventually end-to-end. Various research projects within the 5G PPP programme are working to optimize latency-critical use cases and are developing solutions to reduce delay at user plane and at control plane in specific vertical application contexts. In light of the need for a coherent benchmarking framework, this paper briefly introduces latency definitions and use cases evaluation scenarios from four 5G PPP Phase 2 projects: NGPaaS, 5GCity, 5G ESSENSE and MATILDA. Teams from these project are working towards achieving a common understanding of latency key performance indicators (KPI), trying to converge on common definitions and measurement methodologies to be applied in different vertical use cases and diverse system architectures.**

***Index Terms*— 5G, end to end latency, Key Performance Indicators**

I. INTRODUCTION

Key performance indicators (KPI) measure the quality of system's or organisations performance and are used to monitor performance and level of achievement of wider operational goals.

Many standard organizations have worked towards defining 5G performance KPIs and mechanisms to measure related metrics: recommendations have been produced by International Telecommunication Union-Radio communication Sector (ITU-R), 3rd Generation Partnership Project (3GPP), and Next Generation Mobile Networks (NGMN) Alliance. Overall, these organizations have indicated the advances of 5G systems in a number of improvements compared to previous generation mobile network, as detailed in [1]. For latency, there is a general agreement on trying to achieve 5 times less end-to-end latency reaching delays $\leq$ 5 ms.

A study commissioned by the European Commission in 2015 to derive a framework for monitoring the impact of 5G public private partnership [2] and the associated key performance indicators (KPIs) [3] recently showed how various 5G PPP projects generated a number of specialized KPIs, each responding to the specific application scenarios and 5G sub-system architectures. Based on this scattered operational context, a hard work was needed to reconcile various performance metrics and goals into to a short list of 29 KPIs categorised into three groups: i) 7 Operating KPIs for the program activities; ii) *12 Performance KPIs*, to examine technical capability improvements; iii) *10 Societal KPIs* to assess longer-term outcomes and impacts from 5G technical capabilities and specifically related to 5G PPP projects. Among the others, latency is a key 5G performance factor, and the various elements of the latency assessment (e.g. user plane latency, control plane latency, transmission delays is backhaul or transport, etc.) have been grouped into a parent KPI named end-to-end latency with the overall target goal to reduce it 5 times in comparison with 4g. End-to-end latency which represents the maximum tolerable elapsed time from the instant a data packet is generated at the source application to the instant it is received by the destination application.

This paper is a joint work among four research and innovation projects from within 5G PPP Phase 2: namely, [4], 5GCity [5], 5G ESSENSE [6] and MATILDA [7]. The goal of this work is to provide a common understanding of latency which is a Key Performance Indicator (KPI) of 5G communication across different use cases and architectures. In order to meet the requirements from different vertical use cases, the architecture of 5G system shows high heterogeneity to combine and connect the resources from the physical infrastructure to the upper-layer services tailored to the vertical service providers.

The work is motivated by the need to converge towards a common benchmarking methodology. We have started analyzing the different latency-critical use cases from the four projects (see Sec. II). Then we progressed with the specific per-project definitions of latency metrics aimed at defining commonalities and measurement methodologies to be adopted with the ultimate goal of understanding the feasibility of a common evaluation framework for latency in 5G (see Sec. III). The paper is a preliminary report on

Corresponding author: Gino Carrozzo, g.carrozzo@nextworks.it

status of discussions, with next planned steps briefly hinted as future work in paper's conclusions in Sec. IV.

II. Latency-critical Use Cases and KPIs

In this section the latency-critical use cases from different projects are presented under their novel architecture or ecosystem.

A. MCPTT Service Provider backed by NGPaaS Operator

The NGPaaS project [4] is working on one scenario of MCPTT (Mission Critical Push to Talk), which occurs in a very large factory that has been set on fire: firemen arrived at the site to extinguish the fire, and are divided into several groups to fight against the fire and evacuate people with the support of intensive voice communication between each other and across groups to synchronize 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 virtualized network functions (VNFs) to setup 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 equipped with the infrastructures as an edge data centre with the capability to host User Plane (UP) components of Core Network (CN), components of radio access network (RAN) and MCPTT Apps. The Control Plane (CP) components of Core Network are hosted in a central data centre. Except the MCPTT Apps, all the other components are provided and maintained by NGPaaS operator as a PaaS.

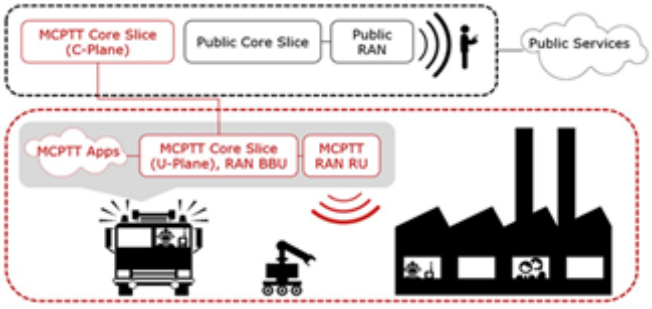

Fig. 1. Architecture of the MCPTT Use Case in NGPaaS

In the MCPTT use case, the PaaS needs to support very low latency ($\leq$at 1 ms for mission critical applications, and $\leq$ 20ms for interactive applications), which will require raising the bar in terms of performance, especially at the RAN level. Along with the requirement of URLLC, the PaaS also faces the challenges to support multiple heterogeneity: 1) workloads of both common IT applications and highly performance-sensitive Telco functions, 2) cloud deployed at both central and edge locations, and 3) cloud infrastructures with different hardware. These requirements and challenges are the motivation to enable the NGPaaS features, which include modularity, build-to-order design principle and various Telco-grade enhancements.

B. Public Safety Use Case of 5G ESSENCE

The 5G ESSENCE project [6] has identified the Public Safety Use Case as one of the most representative examples for URLLC services in multi-tenancy contexts. Two specific scenarios are addressed: Mission Critical Push-To-Talk (MCPTT), and Mission Critical Messaging and Localization (MCML), where a network operator provides an IaaS to the different Mission Critical Organizations (Police, Firemen, Medical System). The process of providing the MCPTT those services in 5G ESSENCE can be summarized as follows:

1) 5G ESSENCE infrastructure operator provides the required network slices to different tenants (Mission Critical Organizations) with its respective SLAs;
2) Allocation of Quality of Service features of each slice is guaranteed by the cSD-RAN controller in accordance with the cloud resources already 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.
3) In case of a damaged ICT infrastructure, 5G ESSENCE Edge Cloud infrastructure maintains the service operation terminating both the control plane and the data plane in the edge deploying VNFs local core functions.

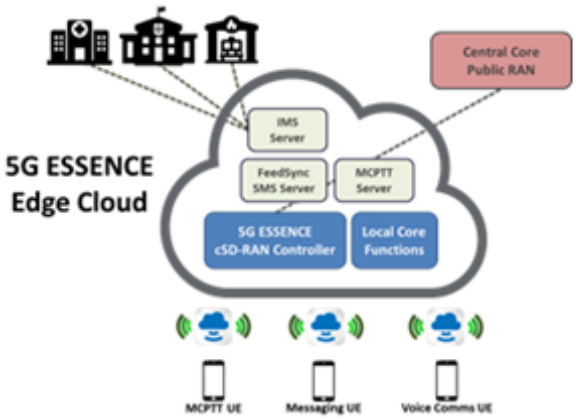

Fig. 2. Edge Cloud of 5G ESSENCE Public Safety Use Case

As depicted in Fig. 2, the 5G ESSENCE Edge Cloud can manage low latency services deployed in the network edge, being able to route to the different Mission Critical Organizations the messaging, data and voice comms 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. Critical to the execution of this server

is the response time in Mission Critical services which is ti be reduced with respect to 4G networks, e.g. with use of edge computing, as well as the decoupling of control and user planes that operate across multiple Edge DCs.

C. 5GCity Media Use Cases for Smart Cities

The 5GCity project [5] is working on the evaluation of 5G Network Slicing and Service Orchestration technologies through media use cases executed in live pilots deployed in the cities of Barcelona (ES), Bristol (UK) and Lucca (IT). The media industry use cases are particularly relevant to the Smart City environment, due to the increasing diffusion among citizens of UHD streaming and immersive media services. In the UHD video distribution and immersive services use case (see Fig. 3) challenging 5G KPIs exist and, in particular, it is critical the available throughput offered to mobile terminals as well as the end-to-end latency (for the parts related to the immersive Augmented Reality service).

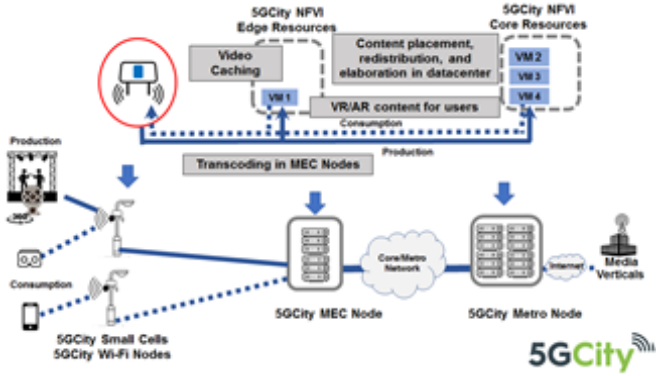

Fig. 3. UHD streaming and immersive services in 5GCity

In this use case the service typology for media distribution and immersive experience is built on-demand and the media consumption takes place while on the move, in mobility across the city areas covered by the 5GCity network. Various types of devices are used (e.g. smartphones, tablets, and virtual reality devices) and various sections of the virtualized network infrastructure are dynamically configured to provision the service (i.e. media servers in 5GCity metro nodes/datacenters, edge computing nodes for local transcoding and video caching, far edge computing for RAN virtualization) . When the immersive aspects come into play, additional contents need to be automatically retrieved from the media server/libraries at datacenter back-end or in edge caches, in the form of 2D video, panoramic video and 3D models to augment the reality in which the user is immersed. In this case, low end to end latency can allow high responsiveness of the immersive application functions. Quantitative KPI targets for this use case are Data Plane Delay (max 10 ms), Control Plane Delay (max 20 ms), coupled with User Experienced Data Rate (100 Mbps DL / 50 Mbps UL).

D. Mission Critical Data in Disaster Relief (MC-DDR) operation in MATILDA

Catastrophic events, such as earthquakes, focus community's attention on the need for powerful and resilient emergency communication networks. 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 MC-DRR scenario in the MATILDA project [7] makes use of this capability to deliver a suite of low-latency services and applications on top of a 5G telecom 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. For such PPDR services to operate reliably and with high survivability and availability, support of assured communications and improved service provisioning intelligence are required, even under extreme conditions and under the assumption of utilizing a distributed 5G network with distributed service intelligence. To support MC-DDR use case, the following actors and stakeholder are part of the MATILDA based 5G emergency ecosystem:

- BB-PPDR network operator provides and operates MATILDA telecom 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.

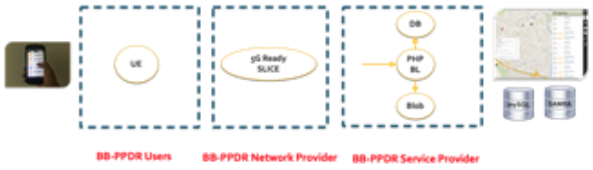

Fig. 4. MC-DDR application for real-time intervention monitoring (iMON) in MATILDA

Beside deployment automation, high availability, resilience and system flexibility, the targeted KPIs for MC-DDR use case include low latency capabilities of 5G user and control plane. 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. Towards a Common Evaluation Framework for Latency in 5G

The system architectures followed by the four projects follow the key principles of the 3GPP TS 23.501, i.e. they separate User Plane functions from Control Plane functions. As defined by the 3GPP standard, the 5G system consists of 5G Core Network, 5G Access Network and UE. Thus, a

common framework for measuring latency can be defined as described in [8] and depicted in Fig. 5, where latency can be measured in the radio segment (T-Radio), in backhaul (T-Backhaul), in edge/core (T-Core) and IP network & cloud (T-Transport).

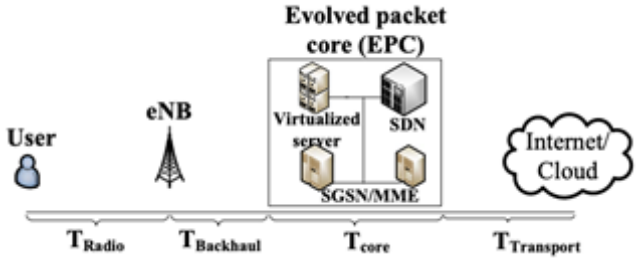

Fig. 5. Contribution of various delay segments to end-to-end latency

For NGPaaS, *Latency* refers to the latency of packets sent from a client to a server based on the RAN Service running on top of Kubernetes with NUMA aware CPU pinning. Network tools will be used to accurately measure the latency of a user flow. The CPU load of the server hosting the RAN components (RCC, RRU) will be progressively increased to measure its influence over quality of 5G PaaS connectivity service.

For 5G ESSENCE and MATILDA, *Latency* refers to packet round-trip time delay on the connected UE. Measurement reference packet should be sent and its response must be received by the same UE device which is able to communicate with the server component. Periodic ICMP/PING measurements from the UE to application components are executed and results are stored in Prometheus platform and reported to the MATILDA orchestrator.

For 5GCity, *Data Plane Delay* and *Control Plane Delay* are, respectively, end-to-end metrics between the UE and a Media Application Server which can be located in edge/core or in the transport network; therefore, end/to end latency refers to the chain of segments shown in Fig. 5. The measure is implemented through RTD test between UE and the media application server. Intermediate delay measures can also be retrieved for the intermediate network sections at edge (where the EPC can be deployed) and in various layers of T-Transport (edge, metro, datacenter).

IV. CONCLUSIONS AND FUTURE WORK

This paper has briefly introduced the approaches followed by four 5G PPP Phase-2 projects (i.e. NGPaaS, 5GCity, 5G ESSENSE and MATILDA) to latency performance evaluation in specific URLLC use cases. Researchers from these project are working towards achieving a common understanding of latency KPI, sharing measurement methodologies and discussing possibilities to adopt a coherent common approach to evaluation for different vertical use cases and diverse system architectures.

Future work will consist of sharing the technical approaches to reduce the latency in specific segments of the network as per Fig. 5, consolidate individually measurement strategies and progress together on this KPI evaluation.

ACKNOWLEDGMENT

This work is supported by 4 H2020 projects which have received funding from the EC within the H2020 research and innovation program for Phase-2 of 5G Infrastructure Public-Private Partnership: NGPaaS (grant 761557), 5G ESSENCE (grant 761592), 5GCity (grant 761508), MATILDA (grant 761898).

REFERENCES

[1] 5G-PPP, ERTICO, EFFRA, EUTC, NEM, CONTINUA and Networld2020 ETP, "5G empowering vertical industries, Available online: http://tiny.cc/1nv25y," Tech. Rep., Feb. 2016.
[2] The 5g infrastrcture public private partnership. [Online]. Available: http://www.5g-ppp.eu/
[3] Tech4i2, Trinity College Dublin, "Framework for monitoring the impact of 5G Public Private Partnership and the associated Key Performance Indicators (KPIs), SMART 2015/0013, Doi:10.2759/79440," Tech. Rep., Sept. 2017.
[4] The ngpaas project website. [Online]. Available: http://www.ngpaas.eu/
[5] The 5gcity project website. [Online]. Available: http://www.5gcity.eu/
[6] The 5g essence project website. [Online]. Available: http://www.5g-essence-h2020.eu/
[7] The matilda project website. [Online]. Available: http://www.matilda-5g.eu/
[8] 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. [Online]. Available: http://arxiv.org/abs/1708.02562