Key Enablers to Deliver Latency-as-a-Service in 5G Networks

Rekha M Das*, Sree Lekshmi S*, Seshaiah Ponnekanti*, Milan Paunovic§

***Abstract* — In the early 4G deployments, mobile communication providers have been focusing on connectivity technologies to meet the coverage and capacity needs of the users. With the advent of 5G, however, new value-added services including Augmented Reality/Virtual Reality (AR/VR), Mission Critical (MC) applications need to be supported. Open Radio Access Network (RAN) together with network slicing technology has been envisioned as one of the important solutions to realize Ultra Reliable Low Latency Communication (URLLC) services in 5G. The paper explores a novel framework and its under-pinning key enablers to deliver the ultra-low latency specifications for various 5G verticals including Smart Cities, Industrial Internet of Things (IIoT) and MC Services.**

Keywords* — *5G, Open RAN, URLLC, Network Slicing, Latency.

I. Introduction

THE cellular networks have been evolving rapidly in terms of architecture and services offered to end users. 5G is the next generation cellular technology standard developed to support diverse use cases such as Enhanced Mobile Broadband, Massive Machine Type Communication and URLLC. The architecture of the 5G network is shown in Fig. 1 and the details of the 5G Service Based Architecture (SBA) can be found at 3rd Generation Partnership Project (3GPP) [1].

URLLC is one of the key services in 5G which supports latency sensitive services such as remote surgery, driverless cars etc. In order to address these requirements, novel Radio Access Network (RAN) framework is noted to be crucial. Traditional RAN architectures are monolithic in nature consisting of proprietary hardware and software components. This framework is lacking the requisite flexibility and hampers speed of deployment. Because of this, Service Providers (SPs) are moving towards Open RAN framework towards agile, flexible and scalable network infrastructure. In typical network deployments, RAN and core network normally contribute to 80/20% split towards the deployment costs.

This becomes possible by utilizing the principles of virtualized edge infrastructure through Network Function Virtualization (NFV), Software Defined Networking (SDN). Both SDN and NFV facilitate creation of logical partitioning of networks by utilising the Network Slicing framework. These technological developments pave the way to implement novel service delivery framework in 5G. By suitably adopting the above trends, URLLC services can be enabled through the 5G SBA fabric. This becomes integral to Service Provider digital platforms. Platform approach contributes to novel business models like Latency-as-a-Service to support new services in the 5G era.

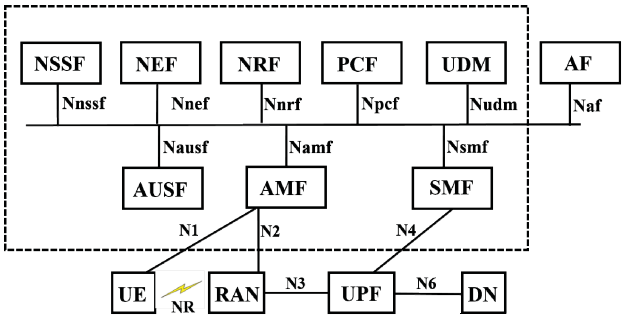

Fig. 1. 5G Service Based Architecture [1]

The paper is organized as follows. Section I outlines introduction to the technology background to the research elements in this paper. Section II discusses some of the key enabling technologies in 5G to realize low end to end latency. Section III describes Latency-as-a-Service framework in the 5G network. Section IV concludes the paper.

II. 5G Key Enablers to Realise Low Latency

The section describes 5G key enablers that supports ultra-low latency applications.

A. Mobile Edge Computing (MEC) based NFV

NFV splits hardware and software resources making network operations as Virtual Network Functions (VNFs) [2]. VNFs provided by Virtual Machines (VMs) can be dynamically instantiated in the network as per the service requirements. NFV acts as a vital enabler of network slicing feature.

MEC enables cloud computational abilities at the edge of the network which is very close to the end user [3]. This helps in reducing the end to end latency when compared to the cloud-based services. The notion of edge corresponds to 5G base station or gNodeB or mini Data Centers (DCs) in the RAN.

MEC framework has been defined by European Telecommunications Standards Institute (ETSI-MEC).

*Amrita Center for Wireless Networks & Applications (Amrita WNA), Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham, India. E-mail address: rekhadasm@am.students.amrita.edu, sslekshmi@am.amrita.edu, seshaiahp@am.amrita.edu

§Mobycore, Serbia - Majevička 2e - Zemun 11080 – Belgrade. E-mail address: milan.paunovic@mobycore.com

978-1-7281-4790-1/19/$31.00 ©2019 IEEE

MEC framework brings closed-loop automation to the edge of the network. The paradigm shift towards edge-based design is considered as one of the main enablers for latency sensitive services in 5G. This framework supports applications and local content providers to effectively utilise edge infrastructure. MEC platforms can run applications at the edge offering secure environment and providing services through RESTful Application Programming Interfaces (APIs). MEC point-of-presence can bring significant advantages like microservices instantiating different applications to form or provide services.

B. Open-RAN

Traditional RAN consists of Remote Radio Head (RRH) and Baseband Unit (BBU) colocated at the cell site. This infrastructure lacks flexibility due to vendor lock-in and use of proprietary hardware and software resources. Open RAN is an innovative solution that can resolve vendor lock-in and confinement to the proprietary resources. Open RAN infrastructure can be built by any vendors, which can interoperate with any other devices or transmission networks attached to it. It builds a modular base station software stack that operates on a Common Off-The –Shelf (COTS) hardware. This allows operators to assemble BBU and RRH from any vendor and put together to form a network.

5G RAN consists of gNodeBs connected to the Core Network (CN). In Open RAN infrastructure, gNodeB comprises three functional modules [4] as shown in Fig.2:

- Radio Unit (RU) - antennas close to the end user
- Distributed Unit (DU) - edge DCs on campus, enterprises, malls or city centers
- Centralised Unit (CU) - bigger DCs (Aggregation points)

Overall RAN network architecture is a combination of wired (Backhaul), wireless and optical interfaces.

- Fronthaul(F2) - Interface between RU and DU
- Midhaul(F1) - Interface between DU and CU
- Backhaul - Interface between CU and CN

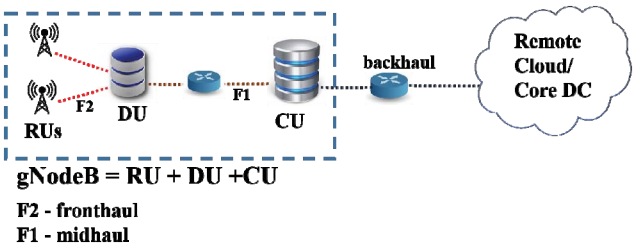

Fig. 2. Functional Modules in Open-RAN

Some deployment scenarios in 5G may utilize enhanced Common Public Radio Interface (eCPRI) in the fronthaul. The interface (F1) between DU and CU is expected to be interoperable across vendors. To cope with the challenges in the RAN such as higher backhaul requirements, protocol splits have been proposed between CUs and DUs. The gNodeB protocol split option is shown in Fig.3. CU can be separated into CU-User Plane (CU-UP) and CU- Control Plane (CU-CP), both of them connects to the DU over F1-U and F1-C interfaces respectively. CU-DU interface is a higher layer split which is more tolerant to delay. The RU-DU interface is a lower layer split which is more latency sensitive and demanding more bandwidth. These RUs, DUs and CUs can be deployed at locations such as cell sites, rooftops and street cabinets depending on the application and network topology.

The mobile traffic is increasing rapidly and varying dynamically, implementing a fixed protocol split may not be a feasible solution. Several criteria are chosen to select appropriate protocol splits. With respect to the traffic variations, protocol split needs to be chosen dynamically.

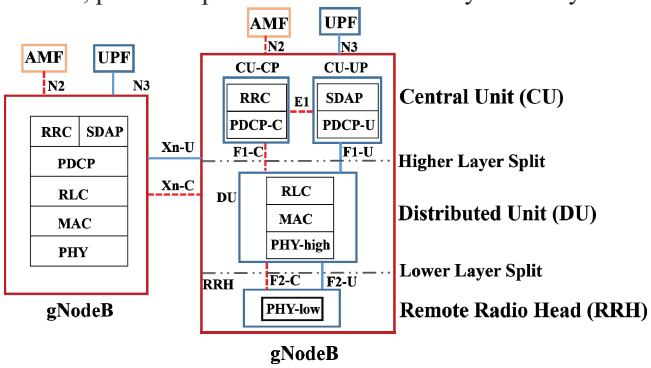

Fig. 3. gNodeB Protocol Split in 5G [5]

The choice between the lower layer and higher layer splits is one of the critical decisions in 5G RAN.

For different regions (rural/sub-urban/urban), different split models can be used. In the urban scenario, where high speed fiber optical connection is available, the following split can be utilised. The units hosting the various protocol stack are as below:

- RU - Low PHY
- DU - High PHY, MAC, RLC
- CU - PDCP

However, for rural area, the entire protocol stack needs to be implemented in the RU itself and no protocol split will be feasible. The unit hosting the various protocol stack are as below:

- RU - Low PHY, High PHY, MAC, RLC, PDCP

There are eight possible combinations of functional split options in 5G as shown in Fig.4 [5]. The deployment of functional modules and protocol split varies with different use case requirements as well as network topology.

C. URLLC

URLLC demands radio latency of less than 1ms with ultra-reliability less than 10^{-5} outage and zero interruption in mobility [6],[7]. The main areas of application of URLLC includes robotic surgery, driverless cars, motion control or harbor automation in factories etc. The CP latency, which is the time taken from idle state to start continuous data transfer should be 10 ms. Whereas UP latency, the time taken to transmit an application from the radio protocol layers through the radio interface in both uplink and downlink is 0.5 ms.

5G New Radio (NR), air interface significantly contributes to the stringent latency and reliability requirement of URLLC [6], [7]. The features that enable low-latency in URLLC has been depicted in Fig.5

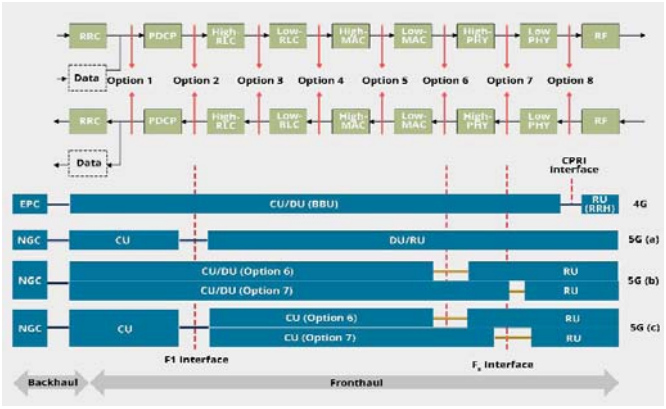

Fig. 4. Protocol Split Options in 4G and 5G [5]

- Hybrid ARQ is moved to the low-PHY which helps in fast retransmissions.
- 5G NR supports scalable numerologies to cater wide range of spectrum, bandwidth, services and deployments. For the data channels, a subcarrier spacing of 15, 30, 60,120 kHz is supported
- Scalable NR slots or mini slots have been used with one slot comprising of 14 symbols. The slot length depends on the subcarrier spacing (SCS) - 1ms for 15 kHz and 0.125ms for 120 kHz SCS. For short transmissions mini slots (2, 4, 7 symbols) can be used and the slots can be aggregated for longer transmissions.
- With Grant free access, radio resources are assigned to the users without any handshake process in prior.

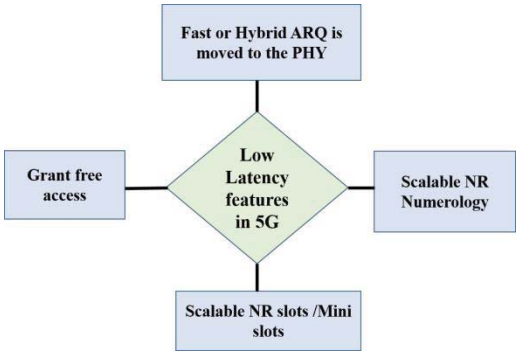

Fig. 5. 5G NR Features that enable low latency

III. Latency-as-a-Service

Network slicing concept provide customised and isolated end-to-end network services to the users. Slicing can logically spin up the network on top of common physical infrastructure by instantiating the required Network Functions (NFs).

Network slice indicates blueprint of the network which describes the type of end devices, appropriate Quality of Service (QoS) to be met for the service type [8]. In order to manage network slices at various stages, the following lifecycle applies to the overall framework:

- **Preparation**: Network Service Descriptor (NSD) describes the network service with essential VNFs which can be executed by the NFV orchestrator.
- **Deployment**: Activation and instantiation of slice and assigning to the users.
- **Run-Time management**: Real-time monitoring of slices to check whether VNFs are required to meet the specific service requirements.
- **Decommissioning**: Deactivation or disabling the slice if not required any more

Slicing as service can enable different verticals [9] like healthcare, public safety, IIoT and automotive industry that belongs to URLLC use case. Multiple tenants can order network slice from SPs or network operators as per the service type and performance requirements. This is called General Slice Template (GST). Upon receiving the slice request from the customer, operators convert GST to Network Slice Template (NST). Single-Network Slice Selection Assistance Information (NSSAI) can uniquely categorise a network slice. NSSAI consists of two components, Slice/Service Type (SST) and Slice Differentiator (SD). Service Orchestrator (SO) [10] can trigger network instantiation process to release the required slice to meet the demands of service type. SO manages end-to-end network service instance as network slice. Slicing facilitates openness of the network to tenants/verticals through Network Exposure Function (NEF) in 5G SBA. NEF securely exposes the services and features provided by 5G Core Network Functions.

The diagram in the Fig.6 below shows the components that could be used within a latency-as-a-service model. An example workflow sequence is given below:

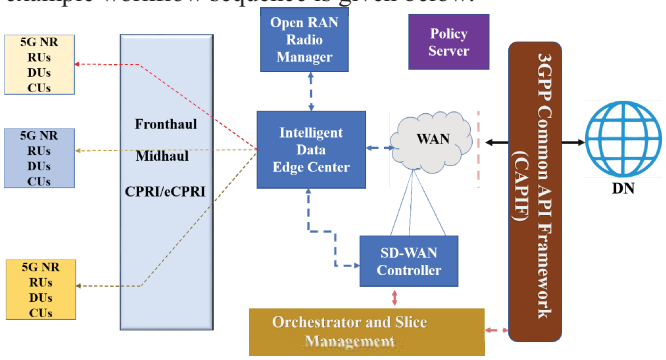

Fig. 6. Framework for Latency-as-a-Service

Step-1: Stakeholders from the vertical side (e.g., Smart City, Tele-health, IIoT, Automotive, Retail or Enterprise) may send a slice request via 3GPP Common API Framework (CAPIF) [11]. This comprises three parts including CAPIF Core Function (CCF), API Exposing Function (AEF) and API Invoker. It has been developed in 3GPP to enable unified Northbound API framework. Both APIs from SPs and 3rd party application developers have been discussed in 3GPP. CAPIF is said to provide a framework to host these diverse set of APIs. Several functions including discovery, registration, authentication, federation are part of CAPIF.

Step-2: The service/slice initiation, maintenance and assurance are part of the management framework. This will be overseen by SO which, in the first instance, liaises with the Policy Control Function (PCF) in the 5G SBA system. The Orchestrator commences provisioning the slice. It

starts with the mobile core control and management of the applicable transport domain controllers (optical, wireless, microwave and ethernet). The objective is to maintain the slice isolation and, at the same time, ensuring that dedicated resources have been assigned for each slice. The transport part of the domain control is governed by Software Defined Wide Area Network (SD-WAN) controllers. In the case, IP/MPLS fabric, all the primary edge routers, aggregators and mobile edge routers will be supervised by the SD-WAN.

Step-3: Intelligent Edge data center exposes its 3GPP and NFV capabilities to applications (normally working edge server and client modes). Edge data center enables application controllers to locally utilize the location and context analytics and maintain the latency low at the edge through intelligent topology selection (RU/DU/CU). Edge computing can be used to implement security policies to harden the network at the edge through edge data center firewall, Deep Packet Inspection (DPI), Identity and Access Management (IAM). This protects the network from malicious upstream traffic from devices like IoT.

Step-4: Open RAN manager supports the distributed multi-vendor virtual radio network functions and hardware radio units. By utilizing the key 5G NR radio interface and toolbox of physical layer technologies, Open RAN manager plays a crucial role in the management of URLLC service. The management part of this module collects the VNFs fault, configuration and performance statics. Unique mapping table is maintained between the slices resource requirements, bearer management and packet delivery at the radio level. This will be addressed together with the fronthaul and midhaul transport capacities to keep the latency minimum at the edge. The following two functionalities make the open RAN manager utmost important to deliver the URLLC slices:

- Dynamic RU selection as per the coordinated measurement reports from all the User Equipment (UE) in the cells/RUs
- Dynamic Protocol split configuration assessment as per the traffic volume in the cells

IV. Summary and Conclusion

In this paper, a novel framework to deliver very low latency services has been introduced to utilize the advances in the 5G technology. The key building blocks of the framework have been introduced to help create the technology context of the paper. Following this, the concepts surrounding network slicing and Open RAN design have been emphasized to show the crucial interplay required to achieve roll out of low latency services in future 5G networks. Finally, service delivery mechanism in terms of the steps and coordination between the enabling technologies in the Latency-as-a-Service framework have been illustrated.

Currently, most of the 5G service launches seemed to be using existing 4G core network infrastructure /Evolved Packet Core (EPC). The radio access, however, is based on 5G NR. In 3GPP terms, this deployment combination of 4G EPC and 5G NR is called Non-Standalone (NSA) mode. It is expected that, Standalone (SA) mode deployments will soon take place with 5G core network (5GC) based on Service Based Architecture (SBA). Recently, Open RAN has received enormous attention due to cost savings to SPs. Several trials have been announced to validate the potential of Open RAN. Based on this demand, as soon as the core networks are upgraded to 5GC, latency-as-a-service model becomes feasible for deployment. Thereafter, it is anticipated to bring about revolutionary benefits in terms of service quality as well as revenues for the industry.

Acknowledgment

The authors deeply thank the support and inspiration provided by the Chancellor of Amrita Vishwa Vidyapeetham, Mata Amritanandamayi Devi (known as “Amma”). This paper has materialized due to her constant guidance. The authors would like to express their gratitude to Dr. Maneesha Vinodini Ramesh for providing immense research support in completing this work.

References

[1] Qiu, Bo-Jun et al. “Service Level Virtualization (SLV): A Preliminary Implementation of 3GPP Service Based Architecture (SBA).” MobiCom (2018).

[2] C. Bouras, A. Kollia and A. Papazois, "SDN & NFV in 5G: Advancements and challenges," 20th Conference on Innovations in Clouds, Internet and Networks (ICIN), Paris, 2017, pp. 107-111.

[3] Y. Yu, "Mobile edge computing towards 5G: Vision, recent progress, and open challenges," in China Communications, vol. 13, no. Supplement2, pp. 89-99, N/A 2016.

[4] Harutyunyan, Davit & Riggio, Roberto, “Flex5G: Flexible Functional Split in 5G Networks: *IEEE Transactions on Network and Service* Management, pp. 1-1. 10.1109/TNSM.2018.2853707, 2018.

[5] Line M. P. Larsen, Aleksandra Checko and Henrik L. Christiansen., “A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks”, IEEE Communications Surveys & Tutorials, Vol. PP, No. 99, 2018.

[6] B. Chang, L. Zhang, L. Li, G. Zhao and Z. Chen, "Optimizing Resource Allocation in URLLC for Real-Time Wireless Control Systems," in IEEE Transactions on Vehicular Technology, vol. 68, no. 9, pp. 8916-8927, Sept. 2019.

[7] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects," in *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124-130, June 2018.

[8] A. Boubendir, F. Guillemin, S. Kerboeuf, B. Orlandi, F. Faucheux and J. Lafragette, "Network Slice Life-Cycle Management Towards Automation," 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Arlington, VA, USA, 2019, pp. 709-711.

[9] D. Raj et al., "Enabling Technologies to realise Smart Mall Concept in 5G Era," 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, 2018, pp. 1-6.

[10] F. Z. Yousaf, V. Sciancalepore, M. Liebsch and X. Costa-Perez, "MANOaaS: A Multi-Tenant NFV MANO for 5G Network Slices," in IEEE Communications Magazine, vol. 57, no. 5, pp. 103-109, May 2019.

[11] 3GPP Specifications 3GPP TS 23.222, TS 33.122 and TS 29.222 (available at http://ww.3gpp.org)