Achieving low-latency communication in future wireless networks: the 5G NORMA approach

Alessandro Colazzo, Riccardo Ferrari, Roberto Lambiase
Azcom Technology srl
Milan, Italy
{alessandro.colazzo, riccardo.ferrari, roberto.lambiase}@azcom.it

***Abstract*—The end-to-end network latency is generally considered by the 5G community a key requirement for future wireless networks, enabling new applications by means of end-to-end figures up to a few ms, which is a target that cannot be achieved by the current 4G technology. 5G Novel Radio Multiservice adaptive network Architecture (5G NORMA) project aims at providing a new network architecture design able to cope with the diverse and stringent 5G KPIs, including network latency. This paper describes the low latency issue from a network architecture perspective, starting from the 3GPP state-of-the-art and then describing the 5G NORMA novelties**

Keywords—latency; 5G technology; networking functions; flexible allocation

I. Introduction

The 5G technology is being designed with many diverse requirements in mind, ultimately coming from several societal and industrial sectors where the new wireless technology will enable a variety of new and diverse applications and use cases that in turn can pave the way to new business models and even introduce new life habits [1], [2]. Such requirements are very demanding for the current wireless technology and not achievable with the mainstream 4G or Wi-Fi systems, due to inherent limitations in the radio interface structure and in the network architecture.

In this paper we analyze the network latency requirement, which is a mandatory constraint for all the applications where real time end-to-end communication is necessary. Several applications which are more demanding in terms of latency are already being investigated and in some cases commercially deployed.

II. principal use cases

The V2V communication requires a latency of maximum 100 ms for safety-critical messages such as the ones exchanged by FCW (Forward Collision Warning) systems. Such figure must be guaranteed also without direct communication between cars, i.e. when the messaging is managed by the network infrastructure. This application, once spread as a mainstream technology, can first have the effect of saving a number of human lives, then can be used for traffic optimization purposes, and then will pave the way for the semi-autonomous and autonomous driving applications [3].

The M2M communication paradigm can be declined in many different applications. As an example, the wireless inter-operability of industrial and robotic devices requires a latency as low as of 10 ms, necessary to have prompt responsiveness of the remote peer system. This technological performance figure will enable new forms of factory management and productive models, generally referred as Smart Factory, where very high efficiency and flexibility are achieved at the same time [4].

The future factory scenario is well complemented by augmented reality applications that are being envisioned in assisted manufacturing and factory management, but also in different uses like assisted driving, augmented cognition, or for critical systems such as for helping in medical interventions [5]. This application foresees a continuous flow of updated data which has to be transformed into visual information and properly positioned according to a number of environmental parameters. The latency required here for the information upgrade is in the range from 15 to 5 ms but we need to take into account the various other sources of delay into the end-to-end processing chain; therefore, to provide that the networking delay is not the bottleneck the network latency is figured out to be as low as 1 ms.

The so called "Tactile Internet" applications involve the tactile sense which is extremely sensitive to the latency performance of the medium. The delay between the given command and the command actuation should be as low as 1 ms. This requirement is valid for applications such remote gaming, remote control of unmanned devices, up to futuristic applications such as the remote surgery. The latency of the associated video application is usually also critical in such cases, since the visual feedback to the controller is given via a remote camera.

III. Latency in 4G LTE

Data about the average latency in 4G commercial networks are easily available and independently measured. All major operators already show average latencies below 100 ms. But such data are average values which can be affected by the network load status and the geographical deployment of the network itself; there is no guarantee for a given latency performance of the network. Moreover, the network architecture does not support any discrimination between a service which needs a given latency performance and another which doesn't. In other words, there is no defined network service that is offered on purpose and that could employ specialized network resources that guarantee the performance figure requested by the end-user application.

The 3GPP consortium is already providing a significant effort in order to enhance the latency performances in 4G LTE standard. Started in Rel'13, TR 36.881 analyzes the latency sources of LTE taking into account control and user plane delays. A number of solutions are proposed, such as grant acquisition process, usage of SPS, handover latency, reduced TTI, all with the objective to improve the network performance for HTTP/TCP applications in order to enhance the user experience and the perceived delay and data rate.

Moreover, there is a dedicated work done with the specific vehicular applications in mind, in order to enable the usage of the already existing infrastructure for the new use cases V2V, V2I, V2P. Rel'14 TR 22.885 gathers the use cases that the standard intends to support and defines the target latency for critical messages as 100 ms and lower [6]. Rel’14 TR 36.885 analyzes the latency requirements of the current 4G technology in order to introduce V2X communication via LTE: the current network performance is analyzed including Uu interface and the sidelink usage, and various improvement items are identified. Such effort is done in order to overcome the fluctuations of the LTE latency due to the network conditions or dimension, and to enable a guaranteed service.

IV. RESEARCH CHALLENGES

In order to obtain such low (up to 1 ms) end-to-end latency figure, it is necessary to analyze the whole chain between the end points in the networks. As depicted in Figure.1, where it is reported a case with a MEC (Mobile Edge Computing) component, the UE data (e.g. a sensor that measures a value) is provided to the embedded system that controls the air interface, then the data passes through the UE protocol stack, the terminal and base station’s physical layer, the base station’s protocol stack, eventually the Mobile Edge Computing processing stack and then goes back through the same reverse chain. In Figure 1, a possible latency budget for the tactile internet is provided. Each element of this communications chain must be optimized for latency.

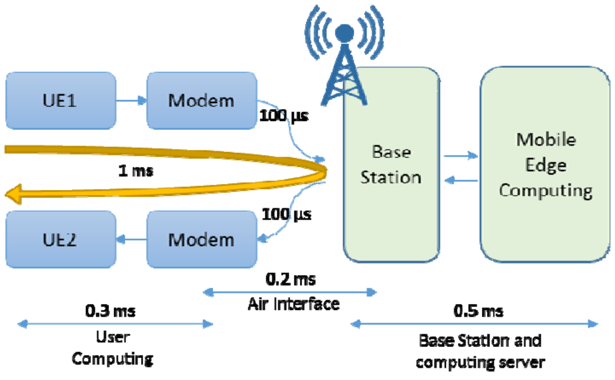

Figure 1. Latency budget

A. *RAN Protocol Stack*

At the air interface level, as reported in [7], the one-way physical transmission should ideally not exceed the duration of 100 µs to achieve a round-trip delay of 1 ms. Then, each packet could possibly have a maximum duration of 33 µs. In LTE every OFDM symbol is roughly 72 µs long, so a new design of the Physical Layer is needed with the 5G system. In [8] the authors suggest the choice of a single-carrier modulation, avoiding so the block-processing of the data which adds extra delay or a tunable OFDM with an adaptive choice of the length of the data block, allowing a TTI lower than 1 ms.

In the U-plane domain, a low latency service would require a very lean protocol stack, where the transition through the various layers is guaranteed to be as fast as possible. Therefore, functions such as MAC and RLC segmentation, concatenation, and especially retransmissions (HARQ and ARQ) must be thought in this direction. The usage of latency-aware and optimized MAC schedulers is also decisive in order to give the necessary priority to the incoming packets [9].

In the C-plane domain most of the RRC procedures have to be redesigned accordingly, since an excessive messages exchange and a non-optimized call context distribution in the network elements would jeopardize the whole budget. First the random access procedure has to be improved by adding a contention free access with dedicated resources [10] or even bringing user data in the PRACH channel itself. With reference to RRC states, special attention has to be put in optimizing the UE RRC states that heavily impact the data transfer setup timings. For example, EU Project METIS-II [11] is investigating the RRC state handling in 5G proposing the introduction of a novel state “Connected Inactive” in addition to “Idle” and “Connected” RRC states. This new state, keeping parts of the RAN context, allows at least a 70% of signaling reduction and a relevant reduction of state transition time. Additional benefits would come from an efficient design of non-access stratum (NAS) and from the integration of NAS and access stratum (AS) since the control signaling could be reduced [12], which is also one objective of the 5G NORMA project.

Moreover, novel methods in higher layers of the protocol stack could reduce end-to-end latency (e.g., avoiding the high overhead of a TCP approach but guaranteeing the desired level of reliability with faster feedbacks).

B. *Backhauling and Network Path*

After protocol stack optimization, the following aspect of the latency budget is related to the back/fronthaul between the radio transmission points and the core network. To reach the 1 ms latency goal, the communication delay related to the physical distance needs to be considered as well. Light, in fact, travels 300 km in 1 ms, then, in theory, a control server that is running a low latency application has to be placed within 150 km from the end user [13]. In reality also the aforementioned delays due to all other sources will contribute thus limiting the maximum distance between the control server and the user which is requiring a low latency service. So the maximum distance should be approximately 15 km [7]. This constraint has led to the introduction of the concept of Mobile Edge Computing (MEC), i.e. having additional processing power near (or into) the base station for local processing at application level.

Beyond geographical proximity, it is worth to consider the optimization of the end-to-end network path, which can be considerably longer than the geographical distance. Reducing the number of involved entities through the network path is a valid technique to diminish latency significantly. In addition, in case of congestion, the involved entities will contribute with tens, hundreds of ms to the latency due to the queues saturation, then reducing the number of hops through the network is even more important. In this sense 5G NORMA is proposing novel network architecture to enable such optimization which will be discussed in section V and VI.

V. 5G NORMA OUTLINE

Considering how diverse and sometimes contradicting the various performance requirements for 5G are, it would be almost impossible to satisfy all of them with a single, fixed network architecture, where we have Radio Access Network (RAN) and Core Network (CN) elements, each one dedicated at running a given functionality or set of.

The key 5G NORMA idea is to have flexible network architecture, where the various networking functions (such as PHY algorithms, scheduling, HARQ, handover, routing...) are run in the RAN or in the CN hardware depending on the service requirement. This baseline idea is allowed by the current ongoing trend of the "cloudification" of the networking hardware, which allows usage of less or not specialized processing elements. The resulted network architecture is therefore cloud based. Such elements can host a variety of processing tasks, even the ones today run on dedicated platforms and accelerators, leveraging virtualization technologies such as Network Function Virtualization (NFV), Software Defined Networking (SDN). These last two enable the so called network slicing concept, that according to NGMN [2], a “network slice (5G slice) supports the communication service of a particular connection type with a specific way of handling the C-and U-plane for this service. To this end, a 5G slice is composed of a collection of 5G network functions and specific RAT settings that are combined together for the specific use case or business model” .

The key enablers of the 5G NORMA approach are briefly described in the following, see [14] for more details.

A. Adaptive (De) Composition and Allocation of Mobile Network Functions

This key enabler is based on NFV paradigm, and support the decomposition of network functions (NF), including access and core functions, and the possibility to associate them to network elements (NE) in a flexible and adaptive way on different levels of infrastructure according to the needed performance. These levels are called clouds, and they can range from the closest to the antenna, called EDGE cloud, to the most centralized, called CENTRAL cloud, as depicted in Figure 2.

B. Software Defined Mobile Networking and Orchestration (SDMN+O)

Complimentary to the previous one, this key enabler takes care of the overall control and orchestration of the flexible allocation of functions in the RAN or in the CN clouds, taking into account service requests, operator's policies, sharing of radio and hardware resources. This is therefore the central intelligence that actually takes care of assigning the deployment of the network functions on the available clouds. Moreover in SDMC+O concept also the control plane functions can be placed arbitrarily in the edge cloud or the central cloud.

C. Joint Optimization of Mobile Access and Core

This third key enabler aims to overcome the drawback of static allocation or distribution of mobile access and core NFs into specified infrastructure entities or network elements. Leveraging on the two previous concepts the network could be split in different slices, a set of dedicated networking resources (hardware, radio, interfaces) to a dedicated purpose, fulfilling customized Service Level Agreements (SLAs). This is governed by the SDMN+O policies which are ultimately defining, dimensioning and setting up such sets of resources.

The flexibility provided by the three enabling technologies allows the design of mobile network architecture, that natively adapts to different types of services, which have different requirements in terms of mobility, latency, traffic volume, security and power consumption.

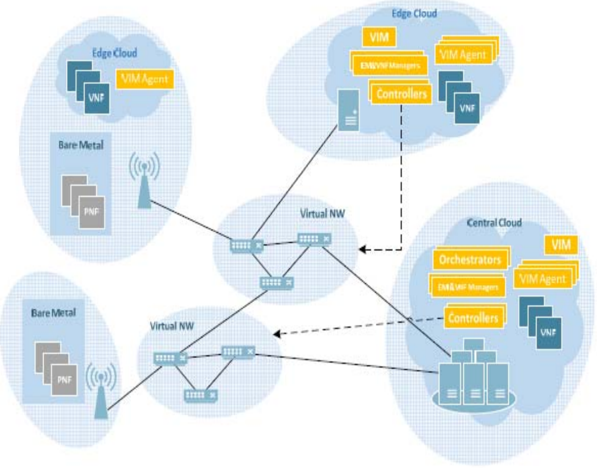

Figure 2. Topological view of the 5G NORMA architecture

VI. LOW-LATENCY NETWORK SLICE

By application of the baseline concepts of 5G NORMA, a low-latency service is provided by a set of NF, belonging to both RAN and CN, where the most critical ones in terms of latency are moved from the central cloud to the edge cloud, in order to avoid backhaul latency and unnecessary hops in the path to the CN.

The intelligence that manages this flexible allocation is the SDMC+O whose primary task is the Service Function Chaining (SFC). A service creation request is mapped to a set of network services (service chain) that in turn is actualized in the instantiation of a suitable network slice configuration; this is done taking into account the service-level requirements such as SLA and key quality indicators. Its final logical output is a service instance that provides the requested service.

The low latency network slice is so instantiated activating the needed NFs in the appropriate clouds. The most critical ones, e.g. routing, are moved from the core network in the edge cloud, close to the user. For even more stringent latency requirements the NFs can be also placed closer to the user, on an edge cloud located on the base station itself. In Figure 3 three different levels of cloud are depicted, depending on the latency required, for three different types of services. The corresponding network slice is kept until the service need ends, afterwards its resources can be reallocated for other purposes.

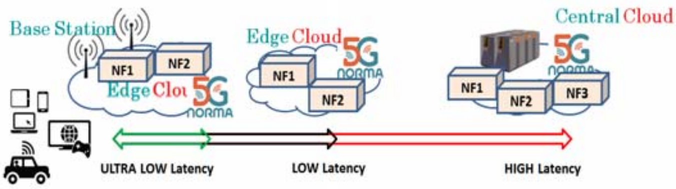

Figure 3. Instantiation of NFs in different clouds depending on the latency requirement of the service

VII. DEMONSTRATING LOW LATENCY BENEFITS

An integral part of the validation process of 5G NORMA project is the realization of Proof of Concept (PoC) demonstrators. PoC demonstrators show that a certain concept or approach is technically feasible and can be implemented with reasonable efforts. 5G NORMA selected three different PoC demonstrators which show the feasibility of function (de)composition and (re)allocation, and of the SDMC approach. The PoC analyzed here is the hardware demo focusing on the function decomposition and relocation and its impact on the latency figure. It also shows practically to a non-technical attendee the benefit of the low network latency and the kind of applications that can be enabled.

The basic idea is to show the latency impact in driving a scale model rally car using a commercial tablet as the steer, both connected to the LTE eNB, as depicted in Figure 4.

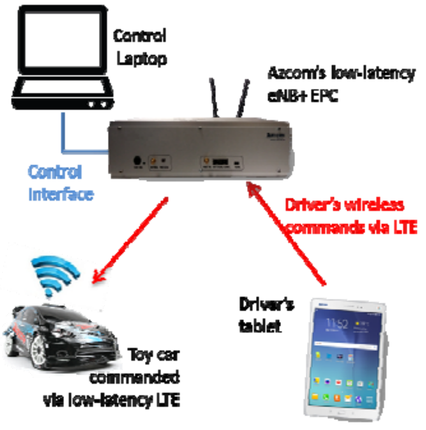
Control Laptop
Azcom's low-latency eNB+ EPC
Control Interface
Driver's wireless commands via LTE
Driver's tablet
Toy car commanded via low-latency LTE

Figure 4. Hardware demonstration of the low latency benefits

Two different situations are demonstrated. The first one mimics a commercially deployed LTE network, in which an average e2e latency of hundreds of ms is experienced. In the second scenario the EPC routing components (S-GW) are moved into the eNB baseband board. Specifically, the S-GW is hosted by the base band SoC as depicted in Figure 5, with the objective to guarantee the lowest latency possible with this HW setup, which in turn would offer to the end user a good driving experience.

The LTE air interface is used in the demo, but packet sizes and RLC/PDCP parameters are set so that the packets go through the protocol stack as quickly as possible. Moreover, after moving the EPC components into the eNB the interface, the S1 interface (now internally managed) will continue to use the GTP-U protocol even if all the components will be running on the same SoC, hence no protocol stack optimization is foreseen.

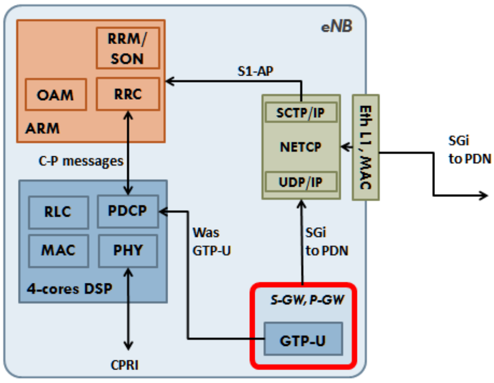

Figure 5. Low latency eNB hosting core network functions

The final application has shown the benefit of the low latency in the controlling the model rally car. In the first scenario the high latency reduces the driving quality, as the command arrives at the model rally car with great delay. In the second scenario the control feeling is very good since the response of the car is immediate. The measured latency in this scenario is between 16 to 18 ms.

VIII. CONCLUSIONS

In this paper the low-latency requirement for future wireless networks has been reviewed starting from the envisioned use cases, that are bringing the necessity of very low figures up to 1 ms or even less. Afterwards an end-to-end latency budget has been proposed, in order to visualize the need of a novel network architecture that, beyond the radio interface, is able to cope with such low end-to-end pass-through figures.

The 5G NORMA architectural approach has been clarified, first in general and then with a specific reference to the low-latency architecture: a dedicated network slice is proposed, which allocates all the needed networking functions as close as it is needed to the radio interface, getting rid of all the core network and backhaul introduced latencies. Beyond the local allocation of the necessary core networking functions, the research challenge is also to design a lean protocol stack, so that in the U-plane domain the packets are able to pass through the protocol stack layers efficiently, and in the C-plane domain the RRC messaging allow very fast data transfer setup and lean call management.

ACKNOWLEDGMENTS

This work has been performed in the framework of the H2020-ICT-2014-2 project 5G NORMA. The authors would like to acknowledge the contributions of their colleagues. This information reflects the consortium's view, but the consortium is not liable for any use that may be made of any of the information contained therein.

REFERENCES

[1] 5G-PPP, "5G Empowering vertical industries", supported by the European Commission, Feb 2015.

[2] NGMN Alliance, "NGMN_5G_White_Paper_V1_0", approved and delivered by the NGMN Board, 17th Feb 2015.

[3] 5G-PPP, "5G-PPP-White-Paper-on-Automotive-Vertical-Sectors", supported by the European Commission, Oct, 2015.

[4] 5GPPP, "5G and the factories of the future – white paper", supported by the European Commission, Oct 2015.

[5] Michael Figl, Christopher Ede, Johann Hummel, Felix Wanschitz, Rudolf Seemann, Rolf Ewers, Helmar Bergmann and Wolfgang Birkfellner, "Latency in Medical Augmented Reality Systems", Proc MICCAI–Augment Environ Med Imaging Comput Aided Surg 1, 2006.

[6] 3GPP TR 22.88, "Study on LTE support for Vehicle to Everything (V2X) services ", Dec 2015.

[7] G. P. Fettweis, "The Tactile Internet: Applications and Challenges," in IEEE Vehicular Technology Magazine, vol. 9, no. 1, pp. 64-70, Mar 2014.

[8] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek and A. Ugolini, "Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," in IEEE Signal Processing Magazine, vol. 31, no. 6, pp. 80-93, Nov. 2014.

[9] I. M. Delgado-Luque et al., "Evaluation of latency-aware scheduling techniques for M2M traffic over LTE," Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, Bucharest, 2012, pp. 989-993.

[10] Aijaz, A., Dohler, M., Aghvami, A. H., Friderikos, V., & Frodigh, M. , "Realizing The Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks", in IEEE Wireless Communications, Accepted for publication, Dec 2015.

[11] S.E. El Ayoubi, M. Boldi, Ö. Bulakci, P. Spapis, M, Schellmann, P. Marsch, M. Säily, J.F. Monserrat, T. Rosowski, G. Zimmermann, I. Da Silva, M. Tesanovic, M. Shariat, A.M. Ibrahim "5G RAN Architecture and Functional Design", METIS II White Paper, [online], Mar 2016

[12] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess and A. Benjebbour, "Design considerations for a 5G network architecture," in IEEE Communications Magazine, vol. 52, no. 11, pp. 65-75, Nov. 2014.

[13] "The Tactile Internet", ITU-T Technology Watch Report , Aug 2014

[14] 5G NORMA, "Deliverable D3.1: Functional Network Architecture and Security Requirements", Feb 2016.