Assessing the impact of computational resources to the Quality of Experience provided by vRANs

Carlos Baena, María Hervás-Gutierrez, Eduardo Baena, Javier Villegas, Raquel Barco and Sergio Fortes, *Member, IEEE,*

***Abstract*—The new generation of mobile networks goes beyond classic radio-focus access networks by adopting the softwareisation of functions that typically relied on special-purpose hardware, now to be deployed in as general as possible purpose platforms. The resulting virtualised Radio Access Network (vRAN) follows the Network Function Virtualisation (NFV) paradigm, enabling a distributed and scalable network architecture. However, this approach increases the complexity of management tasks as computing resources now play an essential role in the network provisioning process. In this sense, this work aims to assess the impact of computational resources on the Quality of Experience (QoE) of video streaming services delivered over vRAN-based networks. The obtained results prove that inadequate resource assignment to vRAN instances leads to degradation of the QoE, even if the allocation of radio resources is adequate for the service.**

***Index Terms*—Mobile networks, 5G, vRAN, computational resources, QoE, video streaming.**

I. Introduction

The arrival of 5G implies a revolution in mobile networks. Compared to previous generations, 5G offers higher simultaneous connection capacity, lower latency, and faster data transmission speeds. These features aim to deliver new user experiences by providing a range of new and enhanced services, such as emergency services, high-quality video streaming, or virtual/augmented reality (VR/AR) services.

In this respect, one of the challenges of 5G is to provide a wide range of services over the same infrastructure. This means a flexible network that can be adapted to any type of use. To this end, 5G networks are being defined on the basis of a service-based architecture (SBA) model. This architectural approach aims to decouple network functions into different services, dividing the network into smaller and modular functional blocks.

SBA relies on technologies such as Software Defined Network (SDN) and Network Function Virtualisation (NFV), which drive the distributed deployment of virtualised network functions (VNFs) across different servers or data centres. This approach has also been extended to the radio access network (RAN), giving rise to virtual RAN (vRAN) paradigm.

This work has been partially funded by Ministerio de Asuntos Económicos y Transformación Digital and European Union - NextGenerationEU within the framework "Recuperación, Transformación y Resiliencia y el Mecanismo de Recuperación y Resiliencia" under the project MAORI. This work has been also supported by Ministerio de Ciencia y Tecnología through grant FPU19/04468 and FPU21/04472.

The authors are with Instituto Universitario de Investigación en Telecomunicación (TELMA), Universidad de Málaga, CEI Andalucía TECH E.T.S.I. Telecomunicación (e-mail: {jcbg, mhervas, ebm, jvc, rbm, sfr}@ic.uma.es).

Corresponding author: Sergio Fortes.

These technologies enable the concept of network slicing, which allows the network infrastructure to be divided into multiple, independent and customisable logical networks, called *slices* [1]. Each slice can be designed and optimised for a specific set of service requirements.

The result is a flexible and efficient End-to-End (E2E) architecture for implementing network services based on user and business needs. In addition, it can be dynamically scaled to meet the changing needs of their users [2].

In this context, traditional network management techniques cannot provide the flexibility required by the 5G paradigm. At the same time, the unbundling (or disaggregation) of the network architecture increases the complexity of managing and orchestrating network elements. In this respect, the radio access network is positioned as one of the critical elements to guarantee optimal network performance, as it is the focus of several references in the literature.

Thus, works such as [3], [4] provide models to characterise the performance of vRANs depending on the level of virtualisation. Moreover, Oliveira et al. [5] evaluate connectivity and capacity to transmit large volumes of data at high speed and low latency over private 5G networks.

Focusing on network slicing, Kalør et al. [6] develop a method to simplify network management in network slicing environments for services with heterogeneous requirements. Similarly, in [7], the authors present an approach based on Machine Learning (ML) that supports the creation of network slices that allow the allocation of radio resources based on the Quality of Experience (QoE) of the service. To this end, they propose a framework that allows the translation of video E2E requirements into network radio configurations.

On the other hand, works such as [8], [9] present solutions for managing and orchestrating the lifecycle of network slices. This includes the preparation, instantiation and configuration of each slice. Thus, these works highlight the importance of considering a relationship between computational resources and radio resources in vRAN environments. In this aspect, there are some works such as [10], where mechanisms are offered for the allocation of PRBs that satisfy certain Quality of Service (QoS) according to the computational resources.

Thus, given the increasing deployment of vRAN-based networks, the study of high-impact factors on the delivery of popular services over this network is of paramount importance. However, to the authors' knowledge, there is no previous work that offers a quantitative analysis of impact of computational resources on the QoE of services offered over vRANs. Here, the present work aims to fill this gap by evaluating the video streaming service in this type of virtualised networks. In this sense, the importance of computational resources in the radio resource allocation process will be highlighted. This is one of the main steps in optimising the QoE of the services offered through the network, as well as the scaling of vRAN entities.

This paper is therefore structured as follows: Section II provides an overview of the RAN in cellular networks, describing its evolution towards 5G NR (*New Radio*) and introducing the concept of vRAN. Next, Section III provides an evaluation of the vRAN paradigm. First, a detailed description of the virtualised network deployment is given, highlighting the most important aspects for the study. Then the results obtained in the virtualised environment are presented. Finally, Section IV presents the conclusions and outlook of the work.

II. The 5G Virtualised Radio Access Network

The RAN is a key part of the mobile network architecture as it is responsible for providing wireless connectivity to users via base stations (BSs). Traditionally, BSs consist of two main components: a radio unit (RU) and a baseband unit (BBU). The RU is the unit responsible for transmitting and receiving radio frequency signals. In this sense, it is responsible for converting radio signals into electrical signals and vice versa. On the other hand, the BBU is responsible for controlling the RU and processing the radio signals for the transport network.

A. 5G New Radio

With the emergence of 5G and the search for a flexible network architecture, the network functions in the BBU are being broken down into less complex units distributed across the network, called functional *splits*. These *splits* are defined by the 3GPP [11] and are numbered from 1 to 8, indicating the separation of BS functions (see Figure 1).

In this sense, the 5G New Radio (NR) architecture decomposes the BBU into two units: DU (*Distributed Unit*) and CU (*Central Unit*). This gives 5G NR more flexibility in terms of infrastructure, as the DU and CU can be distributed according to the requirements of the network architecture into different hardware and/or software units [3]. Thus, both units can be instantiated in the same equipment (which would be a traditional scheme) or in different locations depending on the latency requirements of the network (e.g. CU in a data centre and DU in a location close to the RU). Similarly, the split numbering will indicate which network functions are performed in each of these units. An example of this is shown in Figure 1, where a 7.2 split is shown.

B. RAN virtualisation

Within the NFV paradigm, the vRAN concept aims to decouple all network functions that are implemented on dedicated hardware. The objective is therefore the use of a common hardware platform (General Purpose Processor Platform, GPPP) that hosts the different NFVs. In this sense, in a vRAN architecture, the physical BS (consisting of a DU and a CU) is replaced by a virtual base station (vBS) running on servers distributed in the cloud. This further increases network flexibility and scalability, as well as further optimising hardware resources, as multiple vBS or network entities (i.e. vDU or vCU) can share the same physical infrastructure (see Figure 1).

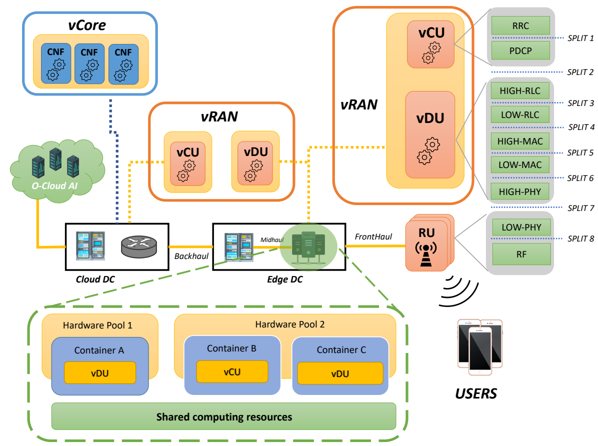

Fig. 1. vRAN distributed scheme

However, this virtualised approach introduces new computational challenges. On the one hand, the distributed location of the elements that make up the vRAN (i.e. RU, vDU and vCU) can introduce latency into the network. Similarly, large amounts of data between vRAN elements can saturate the backhaul and transport network connecting the virtual units, creating a potential bottleneck that introduces latency and packet loss between these units, degrading the service performance.

On the other hand, the management and allocation of each VNF's computing resources plays a fundamental role in the scalability of the vRAN architecture. The processing, memory and storage resources of the physical hardware are distributed across the various virtualised network elements. In this situation, allocating an excessive amount of computing resources to one network element will limit the amount of resources that can be allocated to another network element, thus limiting the scalability of the network. Conversely, insufficient allocation of resources to different network entities can lead to a degradation of vRAN performance, thereby affecting user experience (i.e. QoE).

Optimal management of physical infrastructure resources is therefore essential to enable the efficient deployment of different network elements. This involves taking into account factors external to the network elements, such as the number of allocated radio resources, the number of connected users or the type of carried traffic.

III. vRAN performance assessment

Considering the identified dependency of the vRANs performance depending on computational resources, a complete framework for the assessment of such relation is proposed, as detailed in the next subsection.

A. Evaluation framework

The experimental environment implemented in this work, shown in Figure 2, consists of a 5G NSA (*Non-Stand Alone*)

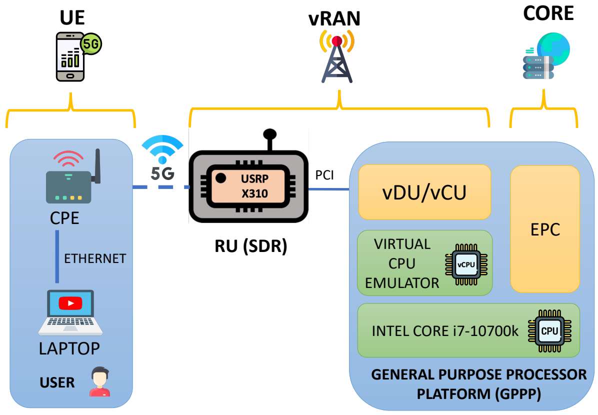

Fig. 2. Network scenario considered

network. It includes a 3GPP Release 17 vRAN, taking into account Split 8.1 (PHY-RF Split).

The RU is deployed using a USRP X310, a Software Defined Radio (SDR) device that allows different radio applications to run on the same hardware platform. It has a Peripheral Component Interconnect express (PCIe) interface that allows connection to any type of GPPP with high data throughput and low latency. The vDU and vCU are implemented as a single virtual unit using the Amarisoft software solution. They are instantiated in a virtual CPU allocation emulator running on an Intel Core i7-10700k at 3.8GHz. This emulator allows the adjustment of the CPU resource usage of the applications instantiated on it.

With this deployment, a set of tests is developed to measure the influence of computational resources on the allocation of radio resources. For this purpose, a radio network slice is created, over which real 4K video traffic from YouTube will be distributed.

For the assessment different number of Physical Resource Blocks (PRBs) are assigned to the slice, while allocating different computational resources to the vDU/vCU through the virtual CPU emulator.

The different resource allocations are evaluated with the demand being generated by an user executing multiple video tests. This user is implemented via a laptop connected to the network through a Customer Premises Equipment (CPE). The laptop will therefore run a set of Python scripts that will allow automatic playback of 4K video via YouTube. In the same way, thanks to web scrapping techniques, various KQIs (Key Quality Indicators) will be obtained [12]. The Mean Opinion Score (MOS) of the service is evaluated using the following equation [13]:

$$MOS = 4.23 - 0.0672 \cdot L_{Ti} - 0.742 \cdot L_{Fs} - 0.105 \cdot L_{Ts} \quad (1)$$

where L_{Ti} is the initial loading time of the video in seconds (related to the initial filling of the buffer), L_{Fs} is the frequency of the video freezing and L_{Ts} is the average duration of the video freezing in seconds. In this way, this formula gives us an evaluation of the QoE of the video streaming service for a fixed resolution, as a function of the initial loading time and the freezing of the image. This gives values on a scale of 0 to 4.23, with higher values corresponding to better MOS and therefore better network performance.

TABLE I
CONSIDER NETWORK PARAMETERS

Parameter	Value
Wireless Technology	5G NR
Bandwidth	40 MHz
SCS (Subcarrier Spacing)	30 KHz
Band	78
Number of slices	1
Slice size (PRBs)	[5,10,15,20,25,30]
virtual CPU allocation	[40,50,60]
Number of users	1
Service	4K Video streaming

B. Evaluation

In order to evaluate the performance of the vRAN deployed in the previously described scenario, a total of 1200 experiments have been launched, corresponding to 50 one-minute plays of 4K video over each network case. For each case, the network slice is configured with different radio resources (in terms of PRBs), while allocating different virtual CPU values. Table I shows a summary of the network configurations considered in this evaluation.

The results are shown in Figure 3, where the different box plots correspond to the evaluation of the video service (via the achieved MOS) in the different network resource allocation cases (PRBs and CPU). In this way, the plots are divided into 4 groups corresponding to the different virtual CPU levels considered. Each group is made up of 6 box plots (grouping 50 experiments each) differentiated by colour, which indicates the size of the slice in terms of PRBs. Similarly, the representation is divided into two parts corresponding to the allocation of part (left) or all (right) of the CPU available in the physical hardware on which the vRAN is running.

The right-hand side presents the impact of PRBs on the QoE of video transmitted over the network. It shows how an increase in the amount of radio resources allocated to the slice translates into an improvement in the delivery of the video service. In this sense, a slice with 5 PRBs is insufficient to deliver 4K video with a MOS higher than 3. Similarly, a slice with 20 PRBs is sufficient to deliver 4K video with the maximum MOS (i.e. 4.23), offering the same performance as in cases where a higher allocation of radio resources is performed.

More interestingly, the left-hand side provides an extended view of this trend, including in this case the impact of computational resources by allocating a certain percentage of the available CPU to the vRAN functions. Here, there is a general degradation in video QoE compared to the previous case (i.e. 100%), which is more pronounced when a more restrictive CPU share is configured. Particularly, with 40% of the CPU allocated, the system is unable to provide good network service. This is reflected in the results by minimum MOS values (i.e. $MOS = 1$). These results are obtained regardless of the allocation of PRBs, indicating a case of

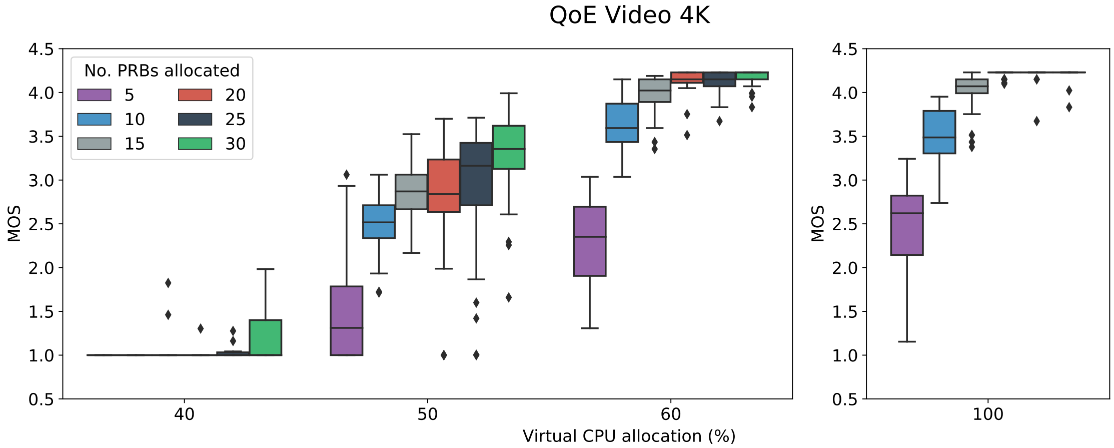

Fig. 3. Service evaluation on network scenarios

network performance degradation due to a deficit in the allocation of computational resources.

When 50% of the CPU is allocated, there is a general improvement for each slice size, with a noticeable improvement in service as more radio resources are allocated. In this respect, the minimum slice size (i.e. 5 PRBs) results in poor MOS values (i.e. $MOS < 2$). These values improve considerably when the allocated radio resources are increased and a correlation with the number of allocated PRBs is observed. However, they are still different from those obtained for the case where the entire CPU is allocated.

Finally, when 60% of the CPU is allocated, the obtained values are very similar to those achieved when the CPU is not fractionated. In this sense, only a subtle dispersion is observed for the case of 20, 25 and 30 PRBs. This shows that by reserving 60% of the CPU, there are no computational problems for offering a 4K video streaming service, allowing the rest of the resources to be used for other VNFs.

IV. Conclusion

This paper has presented an evaluation of the impact of computing resources on application-level QoE. To this end, a quantitative analysis is performed by distributing 4K video over a virtualized network, taking into account the CPU allocated to the vRAN and the assigned radio resources.

The results show how the QoE of the service can be degraded by a deficit in the provision of computing resources. At the same time, it is observed how a less strict CPU allocation allows network performance to be improved by assigning more radio resources.

Finally, it is shown how the proper allocation of resources can provide a good quality of service while freeing up the remaining assets for other VNFs.

Future work will focus on radio architectures with different splits, the evaluation of additional computational parameters such as memory or latency between VNFs, and the impact of network load in virtualised scenarios.

References

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[2] J. S. Panchal, S. Subramanian, and R. Cavatur, "Enabling and Scaling of URLLC Verticals on 5G vRAN Running on COTS Hardware," *IEEE Communications Magazine*, vol. 59, no. 9, pp. 105–111, 2021.

[3] L. M. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146–172, 2018.

[4] L. Diez, C. Hervella, and R. Agüero, "Understanding the Performance of Flexible Functional Split in 5G vRAN Controllers: A Markov Chain-Based Model," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 456–468, 2021.

[5] W. de Oliveira, J. O. R. Batista, T. Novais, S. T. Takashima, L. R. Stange, M. Martucci, C. E. Cugnasca, and G. Bressan, "OpenCare5G: O-RAN in Private Network for Digital Health Applications," *Sensors*, vol. 23, no. 2, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/1047

[6] A. E. Kalør, R. Guillaume, J. J. Nielsen, A. Mueller, and P. Popovski, "Network Slicing in Industry 4.0 Applications: Abstraction Methods and End-to-End Analysis," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 12, pp. 5419–5427, 2018.

[7] C. Baena, S. Fortes, E. Baena, and R. Barco, "Estimation of Video Streaming KQIs for Radio Access Negotiation in Network Slicing Scenarios," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1304–1307, 2020.

[8] I. Afolabi, J. Prados-Garzon, M. Bagaa, T. Taleb, and P. Ameigeiras, "Dynamic resource provisioning of a scalable E2E network slicing orchestration system," *IEEE Transactions on Mobile Computing*, vol. 19, no. 11, pp. 2594–2608, 2019.

[9] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, "Network Slicing Meets Artificial Intelligence: An AI-Based Framework for Slice Management," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 32–38, 2020.

[10] A. Ndikumana, K. K. Nguyen, and M. Cheriet, "Two-level Closed Loops for RAN Slice Resources Management Serving Flying and Ground-based Cars," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2023.

[11] "Study on new radio access technology: Radio access architecture and interfaces," Technical Report 38.801, v14.0.0, 3GPP, Sophia Antipolis, France, 2017.

[12] C. Baena, O. S. Penaherrera-Pulla, L. Camacho, R. Barco, and S. Fortes, "Video Streaming and Cloud Gaming services over 4G and 5G: a complete network and service metrics dataset," *IEEE Communications Magazine*, pp. 1–7, 2023.

[13] M. L. Marí-Altozano, S. Luna-Ramírez, M. Toril, and C. Gijón, "A QoE-Driven Traffic Steering Algorithm for LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 11 271–11 282, 2019.