Improved User Experience by Dynamic Service Handover and Deployment on 5G Network Edge

Jose-Juan Pedreno-Manresa[1], Pouria Sayyad Khodashenas[2], Jose-Luis Izquierdo-Zaragoza[3], and Pablo Pavon-Marino[1]

[1] Universidad Politécnica de Cartagena, Cartagena, Spain

[2] Fundació i2CAT, Barcelona, Spain

[3] Frequentis AG, Vienna, Austria

e-mail: joseluis.izquierdozaragoza@frequentis.com

ABSTRACT

Novel end-to-end services are expected to emerge with the advent of 5G. In order to satisfy QoS/QoE requirements (e.g., lower latency, higher traffic volume and dynamicity, or availability) of those services, major changes on the network architecture are required. Technologies such as SDN and NFV, in conjunction with the edge-computing paradigm facilitate this transition, enabling the presence of cloud-enabled systems on mobile radio access networks. In this paper, we present a novel approach to handle Service Chaining handover, leveraging a joint orchestration between NFV and RAN domains.

Keywords: 5G, NFV, RAN, edge computing, handover, service chaining.

1. INTRODUCTION

Similar to its predecessor 4G, 5G initially started as a direct response to the growing number of mobile devices seeking internet connectivity; however, as it evolved further, we are no longer talking only about connected phones and tablets. The rise of 5G goes along with an explosion of new connected devices such as those related to Internet of Things (IoT), enabling innovative services on different vertical sectors such as eHealth, Industry 4.0 or Autonomous cars. 5G promises not only higher data rates (> 1 Gbps), but also an increase (up to 100-fold) in the number of simultaneous connected devices; as well as lower end-to end latency (< 1 ms) and energy consumption. To fulfil these promises, new antennas, equipment and applications will be required. No matter what the setup looks like, the influx of additional data - which will need to be processed in real-time, will drive the need for edge computing. Recognized by many as the next significant enterprise technology trend, edge computing refers to infrastructure that enables data processing and service provisioning as close as possible to the end user, allowing faster processing and reduced latency, thus, improving customer experience [1].

At the same time, optimized network performance in 5G is an essential factor to meet high user expectations when it comes to low latency, always-on connectivity and innovative services. When 5G powers millions of devices at urban areas, homes, and in workspaces, decisions such as where to send a device's traffic or which equipment is going to process it will need to be made in fractions of a second in order to maintain uptime and provide a superior user experience. The emergence of 5G and edge computing will transform the way enterprises manage their networks and meeting these new challenges. The current paper focuses on the challenges arising in real-life dynamic scenarios (where the amount of devices, traffic and services requested may change rapidly), on top of a network based on very important 5G technologies and enablers (SDN, NFV and small cells). These technologies are important means to meet the 5G promise on ubiquitous coverage, automation and high flexibility and automation, where a joint orchestration of radio and computing resources have already been proved to be essential [2].

2. JOINT RADIO-CLOUD ARCHITECTURE

With the imminent arrival of 5G, all segments of optical networks (transport, metro and access) must not only be upgraded but also undergo radical changes in their architecture to meet all new aforementioned expectations. To accommodate these changes, several new paradigms and architectures have been recently proposed [3][4], being all of them based on bringing the computing capabilities closer to the user, i.e., edge computing. This design paradigm provides a number of advantages, such as reduced latency or easing the traffic load on backbone networks.

One of the proposed solutions for the access segment is called Multi-Tier Cloud-enabled Radio Access Networks (see Fig. 1). This architecture is composed not only by traditional macro cells but also of micro and pico/nano ones, providing additional coverage where needed. All these cells are assumed to be cloud-enabled, that is, they are comprised of IT hardware on top of the radio resources, effectively placing intelligence at the edge of the network. This type of architecture allows the leverage of the Network Function Virtualization (NFV) paradigm, allowing to instantiate Virtual Network Functions (VNFs) such as Network Address Translation (NAT), Firewall, or other virtualized network appliances. To guarantee that end-user services meet the required Quality of Service/Experience (Qos/QoE), the data may be required to traverse and be processed by an ordered

set of VNFs; this is called a Service Chain (SC). Providing total or partial support for SCs directly in the access networks is an essential part of edge computing.

Nonetheless, this joint architecture poses new challenges with respect to orchestration and management. Traditionally, radio and cloud (NFV) domains have been independently managed, each one having its own control plane. In our previous work [3] we showed the presence of certain situations where a high amount of people entering an area (e.g., a football match, a concert or a strike) and the amount of traffic and services requested change very rapidly. In these highly dynamic scenarios where historical traffic or estimations are no longer valid, we demonstrated that a joint orchestration of the two domains is critical for avoiding pathological service blockage situations (e.g., due to insufficient bandwidth from a cell to the Central Office in [2]) and optimizing cloud hardware utilization. Such orchestration would involve and element or agent that communicates with both control planes, coordinating each one's knowledge in terms of users, requested services and associated SCs, the antennas they are connected, their movement, and so on; as well as VNFs/SCs deployed in each cell site, among other information such as session information or QoS/QoE requirements.

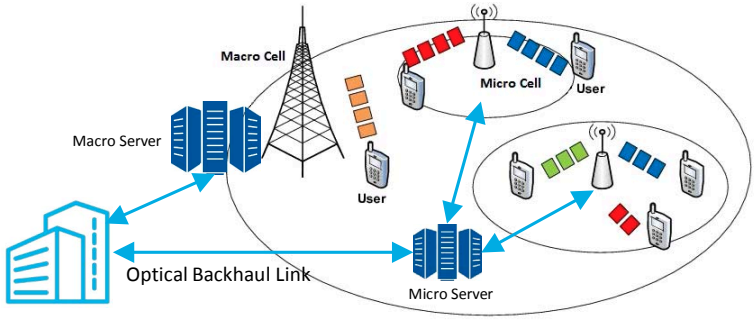
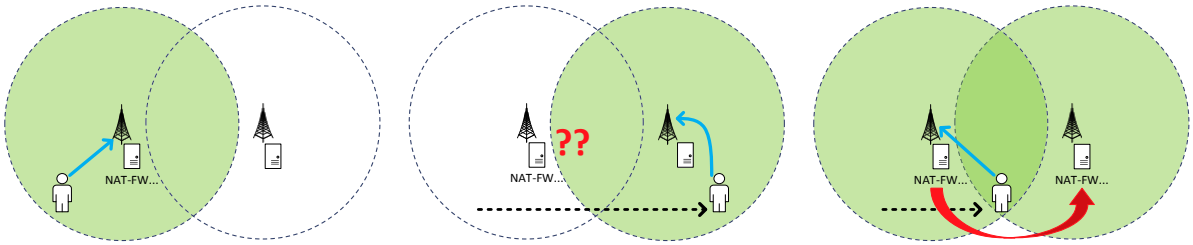


Figure 1. Multi-tier cloud-RAN architecture.

2.1 The NFV Handover Problem

The key use case affected by the need of joint cloud and network resources orchestration is the handover functionality. Handover mechanisms have been recurrently used in Radio Access Networks (RANs) to handle the movement of users across multiples antennas without disrupting phone calls or data usage. In the NFV domain, a handover implies cloud resources reassignments such as data migration or VNFs de/instantiation; however, to the best of our knowledge, no scheme has been yet devised to cover handover actions coordinating radio, network and cloud resources.



(a) Initial scenario (b) Hard handover (c) Soft handover

Figure 2. Service chain handover problem.

Let us suppose a user is actively using a service requiring a specific SC (Fig. 2a), which is partially or totally deployed at the cell serving said user. The user then moves away from the original antenna and under the coverage of another one. The radio handover mechanism takes care of not disrupting the connection with the CO, but it may be the case that the new cell has not yet deployed the necessary VNFs to provide the SC. Deploying new VNFs or migrating the ones from the previous cell, which we refer to as 'hard handover', takes some time (e.g., tens of seconds) and could result in degradation or even interruption of the service (Fig. 2b).

In this paper we study 'soft handover' as a reference benchmark, being this a conservative approach for implementing a non-disruptive NFV services on the network edge. This mechanism works as follows: as soon as a user with an active connection enters the coverage of another cell (Fig. 2c), which is not still managing the connection, the control plane checks whether such cell has deployed the necessary VNFs to support the service in case a real handover is required, proceeding to instantiate new ones if necessary. These resources are pre-allocated until the handover occurs, or they are released in case the user leaves the coverage area without

triggering such handover. Protocol-level details behind this procedure are out of the scope of this paper but would include the refinement of S1-/X2-based mechanisms [5].

Note that this flavor of soft handover may entail huge resource over-provisioning since each cell should preemptively instantiate all the users' services under its coverage area, anticipating a handover that could never happen. Further studies will analyze other scenarios where the number of cells instantiating the service is limited (i.e., using sort of ACTIVE SET mechanisms as in 3G/UMTS). Our interest in this paper is setting a baseline worst-case for resource dimensioning.

3. NFV HANDOVER SIMULATIN AND RESULTS

To test our proposed NFV handover scheme we devised a set of simulations involving a dynamic scenario with a high number of people (approx. 100,000) moving and requesting services in a relatively small area: a football match. We consider an urban area of 1.5 km radius around the Camp Nou stadium in Barcelona (Spain). This area is populated by 56 macro-cells (the exact location of each antenna is provided by [6]). Each cell is assumed to be NFV-enabled, having the necessary hardware resources (i.e., 16 CPU cores, 64 GB of RAM and 1 TB storage) to deploy VNFs, and is connected to the CO via an optical link of 10 Gbps.

At the beginning of the simulation, people are randomly placed at the edge of the simulation area, at specific source points (parking lots, metro stations and so on) and moving towards the stadium, at a random walking speed. Once everyone is inside the stadium, the football match begins (for simplicity, we assume the match duration to be 5 minutes). Upon match ending, people walk back to the starting points, at which time the simulation ends.

Throughout the whole simulation, at each time slot interval (each slot represents 10 seconds), every user has a probability to request a service, based on a binomial distribution taking into account the number of people, the total offered traffic and the service types. Three different services are devised: VoIP, Video Conference, and 5G Service Bundle (e.g., web browsing, 4K video streaming, augmented reality, and so on) with a common holding time of 100 seconds. Each service type has associated a specific Service Chain to be traversed and minimum bandwidth requirements, as shown in Table 1. Hardware requirements to instantiate a particular VNF as well as the maximum number of concurrent operations can be found in in Table 2. During the simulation, the average offered traffic is 15 Gbps and once a user requests a service the probability to be VoIP, Video Conference or 5G Bundle service is 30%, 20% and 50% respectively. Throughout the football match, all services requested inside the stadium are assumed to be handled by a number of micro-cells following a behavior explained in [2][7] (outside of the scope of this work). Lastly, at the end of each time slot, unused VNFs in each cell are de-instantiated to release hardware resources.

Table 1. Service type requirements and associated Service Chains.

Service	Chained VNFs	Bandwidth req.
VoIP	NAT-FW-TM-FW-NAT	250 Kbps
Video Conference	NAT-FW-TM-VOC-IDPS	2 Mbps
5G Service Bundle	NAT-FW-TM-WOC-IDPS	4 Mbps

Table 2. Number of concurrent operations and hardware requirements per VNF.

VNF	# of concurrent operations	Hardware req.
IDPS	2500	CPU: 2 cores, RAM: 2 GB, HDD: 10 GB
FW	2500	CPU: 2 cores, RAM: 3 GB, HDD: 5 GB
NAT	3000	CPU: 1 core, RAM: 1 GB, HDD: 2 GB
TM	2500	CPU: 1 core, RAM: 3 GB, HDD: 2 GB
VOC	1000	CPU: 2 cores, RAM: 2 GB, HDD: 20 GB
WOC	1500	CPU: 1 core, RAM: 2 GB, HDD: 10 GB

Two sets of simulation are performed. In the first one, no dedicated mechanism for NFV handover is assumed, thus hard handover applies; each user requesting a service is handled by the nearest cell which will try to use an existing SC or deploy new VNFs, if needed. When the user has actively moved from one cell onto the coverage of another, and thus the service is handled to the new cell, the same process is repeated: the service is allocated into an existing SC, or new VNFs are instantiated when needed.

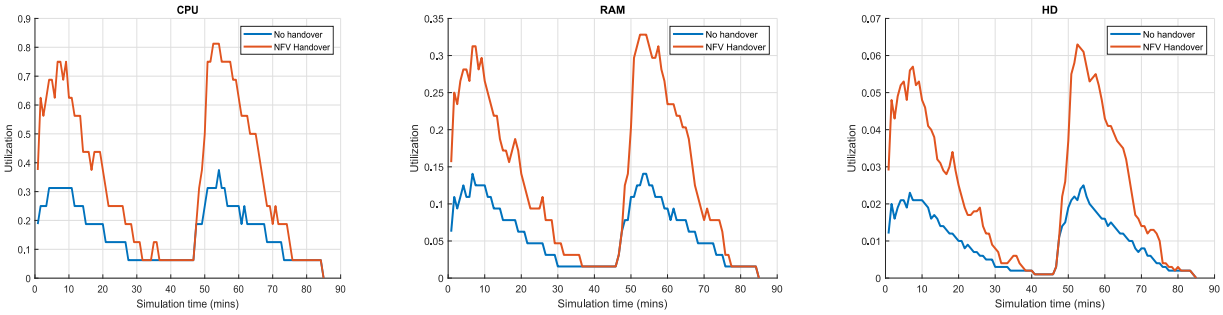
In the second set of simulations, we implement the soft handover mechanism for proper NFV operation. When a user with an active service is under the coverage of another cell (even if it is not handling the service), said cell will assure that enough resources are allocated and available to carry the service in case it is handled to it, preemptively deploying VNFs if needed.

The average hardware utilization (i.e., CPU, RAM and HDD) of each cell throughout the simulation is shown in Fig. 3, for both aforementioned strategies. As expected, using an NFV handover scheme causes much higher utilization due to the pre-allocation of resources, even with no assurance that they will be used. This increase of utilization (up to 3 times), in any case reaches 100%; although at first glance it may unacceptable, we can consider it a compromise considering the worst case no-handover scenario.

Despite the promising results, there is a motivation for further research in handover mechanism that guarantee a smooth transition between cells, taking into account instantiation times while at the same time making a more efficient resource reservation. As an ongoing work, trade-off solutions with limited number of connected cells per user, exploiting user movement predictions and leveraging a joint radio-cloud orchestration are being studied.

4. CONCLUSIONS

Being proven the benefits of a joint orchestration of radio and cloud resources, we explore in this paper the necessity of a handover mechanism to properly handle NFV resources without interrupting or degrading user service. We present and make a worst-case analysis for a soft-handover procedure involving cloud and network resources pre-allocation.



(a) CPU (b) RAM (c) HDD

Figure 3. Comparison of average hardware utilization.

ACKNOWLEDGMENTS

This work was supported by the European Union under the frameworks of H2020 5G ESSENCE (no. 761592), H2020 METRO-HAUL (no. 761727), the FPU fellowship program of the Spanish Ministry of Education, Culture and Sports (ref. no FPU14/04227), the Spanish MINECO/AEI/FEDER project grant TEC2017-84423-C3-1-P (ONOFRE-2) and the "Programa de Formación del PDI a través de la Movilidad de la UPCT" (PMPDI-2018).

REFERENCES

[1] Y. Chao Hu *et al.*, "Mobile edge computing: A key technology towards 5G," *White Paper*, ETSI, 2015.
[2] J.-J. Pedreno-Manresa, *et al.*, "On the need of joint bandwidth and NFV resource orchestration: A realistic 5G access network use case," *IEEE Communications Letters*, vol. 22, no. 1, Jan. 2018.
[3] A. Rostami, *et al.*, "An end-to-end programmable platform for dynamic service creation in 5G networks," in *Proc. OFC 2017*, Mar. 2017.
[4] A. R. Mishra, *Fundamentals of Network Planning and Optimisation 2G/3G/4G: Evolution to 5G*, 2nd ed., Aug. 2018.
[5] J. Prados-Garzon *et al.*, "Handover implementation in a 5G SDN-based mobile network architecture," in *Proc. PIMRC 2016*, Sep. 2016.
[6] Spanish Ministry of Economy and Enterprise – Information About Radioelectric Installations and Radiation Exposure Levels. [Online] Available: https://geoportal.minetur.gob.es/VCTEL/vcne.do (last accessed: Apr. 2019)
[7] J.-J. Pedreno-Manresa *et al.*, "Dynamic QoS/QoE assurance in realistic NFV-enabled 5G Access Networks," in *Proc. ICTON 2017*, Jul. 2017.