Serrano, P., et al. On the benefits of bringing cloud-awareness to network virtual functions, in 2018 *European Conference on Networks and Communications (EuCNC), June 18-21, Ljubljana, Slovenia*

On the Benefits of Bringing Cloud-Awareness to Network Virtual Functions

Pablo Serrano,* Marco Gramaglia,* Dario Bega,*†
David Gutierrez-Estevez,‡ Gines Garcia-Aviles,*† and Albert Banchs*†
*Universidad Carlos III de Madrid, Spain
†Institute IMDEA Networks, Spain
‡Samsung Electronics R&D Institute, UK

***Abstract*—We are currently observing the softwarization of communication networks, where network functions are translated from monolithic pieces of equipment to programs running over a shared pool of computational, storage, and communication resources. As the amount of this resources might vary over time, in this paper we discuss the potential benefits of introducing resource awareness to softwarized network functions. More specifically, we focus on the case of *computational elasticity*, namely, the ability to endure shortages of computational resources while providing an adequate (although non-ideal) service. We discuss how to enable this ability by re-designing network functions, and illustrate the potential benefits of this approach with a numerical evaluation.**

I. Introduction

5G mobile networks will be characterized by a variety of services imposing a diversity of requirements.[1] To efficiently support this diversity, we need a change of paradigm in the provisioning of Network Functions (NFs), moving from the traditional vision where Physical Network Functions (PNFs) are tightly coupled with the hardware substrate running them, to the new vision where Vertical Network Functions (VNFs) run over instantiations of a general-purpose infrastructure. It is envisioned that this transition will introduce a tremendous improvement in flexibility, adaptability and reconfigurability, similar to the one that happened when transitioning from circuit-based to packet-based networking.

By *softwarizing* the operation of the network, VNFs (e.g., schedulers, databases, gateways) run as software components over a set of shared resources (antennas, links, servers, etc.), and can be dynamically provisioned as needed. This approach indeed improves the flexibility of the network: these monolithic programs run over shared computational resources, allowing, e.g., their re-instantiation on-demand, the reduction of development cycles and easier reconfiguration in general. Still, softwarization poses a number of challenges such as the efficient resource assignment to VNFs. This problem has traditionally been investigated from the network management perspective [4], but a very little effort has been done from the VNF design point of view.

[1]This has been repeatedly argued by now, in a number of position papers such as, e.g., [1], in SDOs such as 3rd Generation Partnership Project (3GPP) [2], and in industry fora such as Next Generation Mobile Networks Alliance (NGMN) [3]

In this paper, we focus on the *computational elasticity* of VNF, which we define as the ability to endure shortages of the computational resources while maintaining an adequate service level. The current protocol stack and its composing blocks are "legacies" of the past, namely, their design approach has not changed since early 3GPP Releases. Indeed, the network functions addressed by current softwarization efforts — e.g., Serving Gateway (S-GW), Packet Data Network Gateway (P-GW), or Radio Access Network (RAN) upper layers — have a practically direct mapping with the ones defined in 3GPP Release 8 [6], [7], which was published almost ten years ago. If these VNFs are not designed to run over shared resources (that is, with virtualization in mind), there is the risk that any variation in these resources would cause service disruption: this potential *sensitivity* of VNFs to e.g. resource shortages, updates in the infrastructure, or container migrations, may preclude their wide use in future networks. Furthermore, given the indivisibility of these pieces of software, the assignment of programs to execution nodes (known as network embedding [5]) has been usually performed with a relatively coarse level of granularity, which hinders an efficient use of the resources.

For these reasons, we need a change of paradigm in the design of the network functions, to efficiently support all the novel features that network softwarization and cloudification bring. In other words, the time is ripe for a new class of *cloud-aware protocol stacks*, which embrace softwarization as the fundamental design criteria.

II. The quest for cloudification

The advantages brought by a cloud-driven VNFs design will fuel the research community. In fact, while researchers have devoted so far just a little attention to solve the problems involved by this approach, we believe that the relative maturity of current software initiatives provides the means for research in this area to bloom. We argue that future, fully softwarized and cloudified mobile networks will necessarily build on *cloud-aware protocol stacks*. We believe that both network management and the resulting overall performance will benefit from *making VNFs aware* of being executed in environments such as virtual machines or containers, running on shared resources. One of the key challenges to implement the above vision is to support an elastic operation, to efficiently cope

with changing input loads while running in an infrastructure of resources that is not over-provisioned.

While having a *cloud-aware protocol stack* will benefit any kind of telecommunication service, this may be particularly relevant for the extreme ones. For example, a mission critical VNF can be optimized to reduce its memory footprint, while low latency services may exploit especially tailored orchestration patterns involving edge computing facilities.

An immediate and appealing advantages of a cloudified network is the possibility of reducing costs, by adapting and re-distributing resources following (and even anticipating) temporal and spatial traffic variations. However, it is also likely that in certain occasions the resource assignment across the cloud cannot cope with the existing traffic due to some peaks of resource demands. This is particularly true for Cloud RAN (C-RAN) deployments, that have to deal with demand loads known to be highly variable. In this scenario, allocating resources based on peak requirements would be highly inefficient, as this design jeopardizes multiplexing gains in particular when cloud resources may be scarce (e.g., a "flash crowd" at an edge cloud): here any temporal shortage might result in a system failure. VNFs, instead, shall efficiently use the resources they are assigned with. Thus, they have to become *elastic*, i.e., adapt their operation when temporal changes in the resources available occur, in the same way they have a long-established manner of dealing with outages such e.g. channel errors. Therefore, to fully exploit the benefits of softwarizing the network operation, the network function design has to take the potential scarcity into account, and be prepared to react accordingly.

The concept of elasticity usually refers to a graceful performance degradation when the spectrum becomes insufficient to serve all users. However, in the framework of a cloudified operation of mobile networks that has to deal with elasticity under resource shortages, we also need to consider other kinds of resources that are native to the cloud environment such as computational, memory, and storage assets available to the containers the VNFs are bound to. In particular, we focus on the case of *computational elasticity*: the ability to endure shortages of computational resources while maintaining an adequate service level (although non-ideal). This has hardly been a problem for traditional network functions, that were designed to run over a given hardware substrate with exclusive access to the resources, and requires the definition of novel interfaces that will provide the amount and type of available cloud resources at a given point in time, just like, e.g., the accessible spectrum is a parameter for a RAN function.

Elasticity has also been considered by non-VNFs cloud operators, but our concept deviates very much from theirs: the time scales involved in RAN functions are significantly more stringent than the ones required by e.g., a Big Data platform or a web server back-end. Another key difference is that resources are way more scattered in our scenario (e.g. they are distributed across the "edge clouds"), which reduces the possibility of damping peaks by aggregating resources.

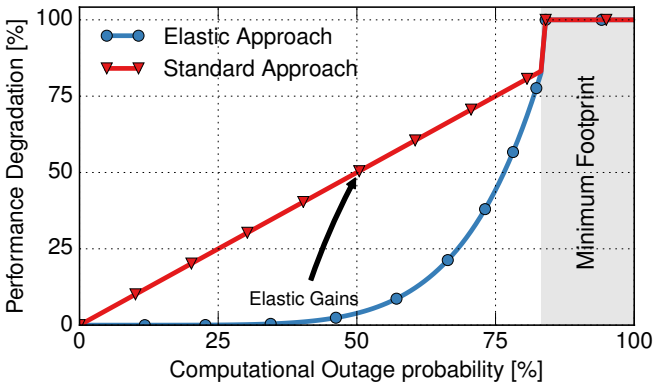

Fig. 1: Graceful performance degradation achieved by elastic computation: performance is not degraded by the same relative amount as resources are reduced.

A. Potential Benefits

To better illustrate the benefits of elasticity in the cloudified mobile network operation context, we first consider the notion of "computational outage" [8], i.e., the unavailability of the required resources to perform the expected operation. In a traditional, non-elastic operation, there is a 1-to-1 mapping between outages and performance loss, as Fig. 1 illustrates: if the resources are not available 20% of the time, there is a 20% performance degradation, as the function is unable to operate under any shortage. In contrast, an elastic design supports a graceful performance degradation, which causes that the VNF still functions under a resource shortage, this resulting in the "gains" illustrated in Fig. 1. Making a protocol stack *cloud-aware* through elastic VNFs requires hence a paradigm shift in their design, moving away from the tight hardware-software co-design that we discussed before, to a flexible operation in which the amount of available resources is an additional parameter.

To fully take advantage of elastic VNFs, a detailed analysis of their operation is required: first, a thorough assessment of the resources consumed during execution, including statistics about temporal variations over time; second, a characterization of the correlations between VNFs operations, to serve as input for the orchestration algorithm, so it could e.g. dynamically assign resources to resilient VNFs and quickly "rescue" them when outages happen.[2]

III. A Practical Scenario

We next discuss the potential benefits of bringing cloud-awareness in a practical scenario. We consider a C-RAN scenario where the scheduling of a number of base stations is done by a central entity. In a traditional approach, the VNF performing frame decoding has to be dimensioned for peak capacity, i.e., all Physical Resource Blockss (PRBs) using the highest Modulation and Coding Scheme (MCS), which

[2]Indeed, the quest for cloudification will end up with novel orchestration algorithms.

corresponds to ideal radio conditions for all users (having a set of users with good channel conditions is a common assumption for schedulers that rely on Opportunistic Scheduling techniques [9]). However, planning for peak capacity not only requires prior knowledge of the users' demand, but also results in resource wastage when mobile traffic falls below this peak.

Thus, let us assume a C-RAN scenario where resources are not over-provisioned, and the scheduling function for the uplink is serving a large enough set of base stations. Under these conditions, it will be likely that during short periods of time a set of users (experiencing good channel quality and hence using high MCS) require more capacity than available, as a higher MCS requires more iterations to be decoded [10]. A non-elastic function would fail to e.g. decode the PRBs, this resulting in an abrupt degradation of performance.

A cloud-aware MCS selection function helps to address this challenge more efficiently. Given that the "disruption" is caused by a relatively large set of users using high MCS, one elastic strategy is to purposely "downgrade" some of the MCS to be used in the PRBs, to find a combination that can be supported by the available computational resources. This version of the function, originally proposed in [11], might have better short-term fairness properties than the previous one, as users are still scheduled but at a lower rate. Also, this might support a more "graceful degradation" in the absence of resources.

REFERENCES

[1] P. Rost *et al.*, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Communications Magazine*, 2014.

[2] 3GPP TR 22.891, "Feasibility Study on New Services and Markets Technology Enablers," 2015.

[3] NGMN Alliance, "5G white paper," 2015. Available Online: http://www.ngmn.org/5g-white-paper.html

[4] A. De La Oliva *et al.*, "Xhaul: toward an integrated fronthaul/backhaul architecture in 5G networks," *IEEE Wireless Communications*, 2015.

[5] A. Fischer *et al.*, "Virtual network embedding: A survey," *IEEE Communications Surveys & Tutorials*, 2013.

[6] E. Dahlman *et al.*, "3G evolution: HSPA and LTE for mobile broadband," *Academic press*, 2010.

[7] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description;" version 8, 2007.

[8] M. Valenti and P. Rost, "The role of computational outage in dense cloud-based centralized radio access networks," in *IEEE GLOBECOM*, 2014.

[9] A. Asadi and V.Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Communications surveys & tutorials*, 2013.

[10] S. Bhaumik *et al.*, "CloudIQ: A framework for processing base stations in a data center," in *ACM MOBICOM*, 2012.

[11] P. Rost *et al.*, "Computationally aware sum-rate optimal scheduling for centralized radio access networks," in *IEEE GLOBECOM*, 2015.