Dynamic Deployment of Virtual Network Functions in Heterogeneous Telco Clouds

Antonio De Domenico, Nicola di Pietro, Ghina Dandachi, and Emilio Calvanese Strinati
CEA-Leti Minatec Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 09, France

***Abstract*—The next generation of cellular networks will leverage on network slicing to deal in a cost efficient way with the service requirements of the vertical markets. With this approach the operator can logically split a single network infrastructure in multiple instances whose virtual network functions (VNFs) are designed for each specific service. This new paradigm requires new functionalities and interfaces to take fully advantage of the pool of computational and communication resources and VNFs shared across multiple network slices. In this paper, we introduce the concept of orchestration-driven elasticity that enables the 3GPP Network Slice Management Function (NSMF) to dynamically adapt the VNF placement in the cloud infrastructure according to the momentary service requirements, and to implement proactive measures that avoid computational outages.**

I. Introduction

The fifth generation (5G) of cellular communication systems currently under design aims to accommodate a wide range of services, use cases, and applications derived by vertical industries, in particular in terms of latency, resilience, coverage, and bandwidth. To achieve this objective, 5G requires a flexible, adaptable, and programmable architecture that leverages on network virtualization and softwarization to create dedicated logical instances of the network, i.e., network slices [1], each one characterized by ad-hoc functionalities and resources.

This novel network architecture needs to integrate innovative mechanisms that jointly adapt the usage of communication and cloud resources to the momentary service requirements. This problem is very challenging: on the one hand, the functions that manage the cloud resources operate in a much slower time scale with respect to the corresponding functionalities used to adjust the communication resource usage. On the other hand, in a virtualized mobile network, there is a tight coupling between the usage of radio and computational resources: e.g., a computational outage may be experienced when allocating more radio resources to increase the throughput of a given slice, without increasing the corresponding cloud resources [2].

One potential solution to such problem is to design management and orchestration functionalities based on the resource elasticity principle. In the context of 5G telco cloud, this principle refers to the ability to adapt in an automatic manner the system configuration to the available resources and network requirements, potentially leading to a graceful performance degradation but avoiding service outages.

This work has been performed within the 5G-MoNArch project, part of the Phase II of the 5th Generation Public Private Partnership (5G PPP) and partially funded by the European Commission within the Horizon 2020 Framework Programme.

The 5th Generation Public Private Partnership (5G-PPP) project 5G MonArch is investigating three levels of elasticity [3]: computational, orchestration-driven, and slice-aware elasticity. These three elasticity principles are highly interconnected and the associated optimization mechanisms have to work jointly by design. In this paper, we will focus on the orchestration-driven elasticity concept by discussing its potential benefits and implementation challenges.

II. Orchestration-Driven Elasticity

Orchestration-driven elasticity focuses on the ability to reallocate NFs within available cloud resources exploiting both central and edge clouds while taking into account the service requirements and the network resource usage, and implementing measures to avoid future computational outages.

Actually, this principle can be implemented in both reactive and proactive ways. Orchestration-driven elasticity can be used to adjust the NFs deployment after a variation of resource usage: for instance, when some resources in the edge cloud are released, time constrained NFs can be moved from a more remote central cloud to the edge, in order to improve the end user performance. To efficiently implement such functionalities, special attention needs to be paid to (i) the trade-off between central and edge clouds and the impact of choosing one location for a given function, and (ii) the coexistence of Mobile Edge Computing (MEC) and network functions in the edge cloud. In contrast to the reactive solutions, by carefully monitoring the cross-slice resource usage trends, the orchestrator can move some NFs from a cloud node to another in order to prevent lack of computational resources. Therefore, orchestration-driven elasticity operates at both cross-slice and intra-slice level.

III. System Model

In our architecture, in contrast to the more classical Centralized Radio Access Network (CRAN), a distributed and heterogeneous cloud infrastructure is considered. This choice enables to move NFs between central and distributed cloud units to allow load management, performance optimization, and adapt the system configuration to the momentary load of the transport network (see fig. 1). Let denote as $\mathcal{N} = \{1, 2, \ldots, N\}$ the set of NFs that compose the 5G network. Each request by a NF $n \in \mathcal{N}$ is characterized by specific computational requirements x_n in terms of number of operations to be processed.

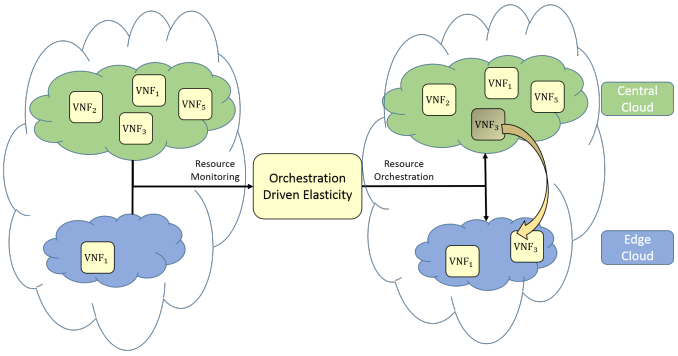

Figure 1: Moving VNFs from the central cloud to the edge cloud through the Orchestration-Driven Elasticity.

Then, we indicate as $\mathcal{I} = \{e, c\}$ the set of available (edge and central) cloud units. Each cloud unit $i \in \mathcal{I}$ is identified by its computational capacity C_i, which represents the maximum number of computation operations per second that it can support.

In the current technology, NFs are characterized by strict inter-dependencies, i.e., signaling and data need to go through neighboring NFs within specific timing constraints, in order to guarantee reliable network operations. The most stringent constraint exists at the lowest level (the physical layer) of the access network protocol stack, i.e., between the first NF that can be centralized and the functions located at the access point. To consider this, we define an additional NF, denoted NF 0, which includes those functions that cannot be virtualized.

In the considered architecture, NFs are executed in a serial flow; the latency experienced by this flow includes both the computational delay l^{C} and the communication delay l^{T} of each NF. The former depends on the computational requirement of a NF and on the amount of computational resources that it receives. The latter is experienced when the two communicating NFs are not located in the same cloud unit, and the flow needs to be transferred through a communication link. To model the communication delay, we consider a transport network technology with rate R^{T} [bps/Km] connecting the edge and the central cloud nodes, and the cloud units with the access network. In addition, we consider that the output data w_n [bits] generated by the NF n, which has to be transferred to its neighboring NFs, linearly depends on the computational requirements x_n, i.e., $w_n = \alpha x_n$, where α [bits per operation] is a fixed system parameter.

IV. Problem Statement and Analysis

Here we characterize and analyze the problem of optimizing the deployment of a set of NFs over the cloud units jointly with the associated resource allocation. To strike a balance between the resources used at the central and edge cloud, we consider a convex price function $P_i(x)\ \forall\ i \in \mathcal{I}$ characterized by a two-part tariff [4]:

$$P_i(x) = c_i^d x^2 - g_i x + c_i^c,$$

where $g_i x$ describes the gain due to the usage of the computational resources x, $c_i^d x^2$ denotes the cost for deploying x computational resources, and c_i^c represents the cloud service connection cost.

Since the central cloud can be located in a suburban or rural area, its connection costs are larger than those for the edge cloud, $c_c^c > c_e^c$; in contrast, we assume that the resource deployment costs are larger for the edge cloud, $c_e^d > c_c^d$; however, the gain provided by the usage of computational resources at the edge cloud is larger than the one provided by the central cloud due to the reduced communication latency, $g_e > g_c$.

According to this model, the joint NF deployment and resource allocation problem can be formulated as follows:

$$\min_{\mathbf{C},\mathbf{a}} \quad \sum_{i\in\mathcal{I}} P_i \left(\sum_{n\in\mathcal{N}} C_{i,n} \right) \tag{1a}$$

$$l_n^{\mathrm{C}} + l_{n,n+1}^{\mathrm{T}} \le l_{n,n+1}^{*} \quad \forall\ n \in \mathcal{N}\backslash\{N\} \tag{1b}$$

$$l_n^{\mathrm{C}} + l_{n,n-1}^{\mathrm{T}} \le l_{n,n-1}^{*} \quad \forall\ n \in \mathcal{N}\backslash\{1\} \tag{1c}$$

$$l_1^{\mathrm{C}} + l_{1,0}^{\mathrm{T}} \le l_{1,0}^{*} \tag{1d}$$

$$a_n \in \{0,1\} \quad \forall\ n \in \mathcal{N}, \tag{1e}$$

where $C_{i,n}$ denotes the amount of computational resources allocated by the node i to the NF n and a_n is a binary deployment variable with $a_n = 1$ if the NF n is executed at the edge cloud and $a_n = 0$ otherwise. Moreover, (1b)-(1d) describe the latency constraints between a NF n and its neighboring NFs $(n+1, n-1)$, with computational delay

$$l_n^{\mathrm{C}} = \frac{a_n x_n}{C_{e,n}} + \frac{(1-a_n)x_n}{C_{c,n}} \quad \forall\ n \in \mathcal{N} \tag{2}$$

and delays related to the communication links

$$l_{n,n-1}^{\mathrm{T}} = \frac{w_n d_{e,c}}{R^{\mathrm{T}}}(a_n(1-a_{n-1}) + a_{n-1}(1-a_n)), \quad \forall\ n \in \mathcal{N}\backslash\{1\} \tag{3}$$

$$l_{n,n+1}^{\mathrm{T}} = \frac{w_n d_{e,c}}{R^{\mathrm{T}}}(a_n(1-a_{n+1}) + a_{n+1}(1-a_n)), \quad \forall\ n \in \mathcal{N}\backslash\{N\} \tag{4}$$

$$l_{1,0}^{\mathrm{T}} = \frac{w_1}{R^{\mathrm{T}}}\left(a_1 d_e + (1-a_1)d_c\right); \tag{5}$$

$d_{e,c}$, d_e, and d_c denote respectively the distances [Km] between the edge and cloud nodes and the between the edge and central clouds and the access network.

Solving the mixed-integer problem described in (1a)-(1d) is very challenging for three main reasons: first of all, the binary variable a_n makes it combinatorial; second, in (2) there is a coupling between the two optimization variables, which makes the problem also non-convex; finally, there are additional couplings between a_n and a_{n-1} in (3) and a_n and a_{n+1} in (4). At this stage, we have identified two possible approaches to solve the joint NF deployment and computational resource allocation problem: the first approach targets to find a near-optimal solution with a low complexity by relaxing the integer variable and reformulating (1a)-(1d) in a convex optimization

problem; the second approach is based on the matching theory [5], a mathematical framework that provides tractable solutions for combinatorial problems related to the matching of players belonging two in two distinct sets. Solving this problem and analyzing the proposed solution is left as a future work.

REFERENCES

[1] Next Generation Mobile Networks (NGMN) Alliance, "5G white paper", Feb. 2015.

[2] P. Rost, S. Talarico, and M. C. Valenti,"The complexity–rate tradeoff of centralized radio access networks," in *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164-6176, Nov. 2015.

[3] D. M. Gutierrez-Estevez *et al.*, "The path towards resource elasticity for 5G network architecture,", in *Proc. Workshop on Flexible and Agile Networks (FlexNets), IEEE WCNC*, Barcelona, Spain, Apr. 2018.

[4] J. K. MacKie-Mason and H. R. Varian, "Pricing congestible network resources," in *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1141-1149, Sep. 1995.

[5] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," in *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52-59, May 2015.