Optimal Placement of User Plane Functions in 5G Networks

Irian Leyva-Pupo, Cristina Cervelló-Pastor, Alejandro Llorens-Carrodeguas

Optimal Placement of User Plane Functions in 5G Networks

Irian Leyva-Pupo[1(✉)], Cristina Cervelló-Pastor[1], and Alejandro Llorens-Carrodeguas[1]

Department of Network Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain.
{irian.leyva, cristina, alejandro.llorens}@entel.upc.edu

Abstract. Because of developments in society and technology, new services and use cases have emerged, such as vehicle-to-everything communication and smart manufacturing. Some of these services have stringent requirements in terms of reliability, bandwidth, and network response time and to meet them, deploying network functions (NFs) closer to users is necessary. Doing so will lead to an increase in costs and the number of NFs. Under such circumstances, the use of optimization strategies for the placement of NFs is crucial to offer Quality of Service (QoS) in a cost-effective manner. In this vein, this paper addresses the User Plane Functions Placement (UPFP) problem in 5G networks. The UPFP is modeled as a Mixed-Integer Linear Programming (MILP) problem aimed at determining the optimal number and location of User Plane Functions (UPFs). Two optimization models are proposed that considered various parameters, such as latency, reliability and user mobility. To evaluate their performance, two services under the Ultra-Reliable and Low-Latency Communication (URLLC) category were selected. The acquired results showcase the effectiveness of our solutions.

Keywords: 5G · User Plane Functions Placement (UPFP) · MILP

1 Introduction

The Fifth Generation (5G) of mobile networks has been envisioned as a system capable of overcoming current network limitations as well as an enabler for the development of industry and society. Among the wide range of service scenarios expected of 5G networks, those that fall under the Ultra-Reliable and Low-Latency Communication (URLLC) category are the most challenging to fulfill because of their strict requirements in terms of reliability and latency.

To this end, many research studies have presented their primary target as an air interface, control and/or user planes design, handover (HO) procedures management, or network functions (NFs) placement. The present paper focuses on the last category, specifically, the placement of the User Plane Functions (UPFs). UPFs are the main NFs within the 5G user plane and play a similar role to that of Serving Gateways (SGWs) and Packet Gateways (PGWs) in Evolved

Packet Core (EPC) networks, with the main difference being that UPFs only perform functions related to the user plane.

In 5G networks, services with high demands on latency and bandwidth require the movement of NFs such as gateways, toward the local or central office data centers (DCs) through a downward shift. This means that the number of gateway nodes (e.g., UPFs) must increase by a factor of 20 to 30 times the original amount [1]. A higher number of UPFs will not only result in an increase in network operator expenditures but also in UPF relocations. The latter occurs because of user mobility when a user attaches to a radio access node served by a UPF that differs from the one of its source access node.

Unnecessary relocations can severely impact users' Quality of Experience (QoE) by incurring additional delays and signaling during handover procedures, thereby leading to the necessity and importance of optimal UPF placement. This enables the stringent requirements of 5G networks to be more effectively coped with while simultaneously reducing capital and operational expenditures.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of selected studies that are related to mobile gateway placement and reliability metrics. Section 3 introduces the 5G user plane reference architecture. Section 4 presents two Mixed-Integer Linear Programming (MILP) models to address the UPF Placement (UPFP) problem. Section 5 evaluates and compares these MILP models as well as present an extensive analysis of their results. Finally, Section 6 concludes the paper and suggests directions for future studies.

2 Related Work

In this section, selected studies related to the placement of mobile gateways and reliability metrics are reviewed.

Taleb and Ksentini [2] asserted the importance of considering gateway relocations in reducing costs as well as their impact on users overall QoE. The authors formulated SGW placement as a service area planning optimization problem aimed at reducing the costs of gateway relocations subject to SGW capacity restrictions. Similarly, in [3], the SGW placement problem was addressed from the perspective of SGW relocations; however, the main objective was not only to minimize relocations but also to minimize the load in SGWs. In [4], the authors proposed an algorithm to place virtual instances of PGWs with the aim of reducing costs while ensuring QoE. To this end, the load assigned to PGWs and their imbalance were optimized; nonetheless, they overlooked service latency requirements and the occurrence of PGW relocations. In [5], the placement of SGWs and PGWs was addressed by considering delay and relocation constraints. In this paper, various algorithms aimed at minimizing SGW relocations and the paths between users and PGWs were presented.

Much of the literature addressing the placement of mobile gateways has focused on specific parameters such as capacity, relocations and latency. However, none of these studies have addressed all of these metrics at once. Moreover, solutions regarding the use of reliability metrics in the placement of mobile

gateways are missing, despite being utilized in a wide variety of studies related to the placement problem. In particular, those papers tackling the placement of Virtual Network Functions (VNFs) and Software Defined Networking (SDN) controllers [6–8] have relied on reliability considerations for their solutions.

Liu et al. [6] jointly addressed the placement of SDN controllers and satellite gateways in a 5G-satellite integrated network. Their main objective was to determine the most reliable locations for SDN controllers and satellite gateways to maximize the average reliability for a given number of controllers and latency constraints. Authors in [7] proposed the Resilient Controller Placement (RCP) which assigns the switches to m resilient levels of SDN controllers to enhance the resilience of the control plane. The RCP was aimed at minimizing the total incurred cost by considering the number of controllers and propagation latency, mainly. Likewise [7], Tanha et al. in [8], proposed assigning switches to r levels of controllers to improve resilience. Their main objective was to minimize the number of deployed controllers subject to resilience levels, latency and capacity requirements. Although their method guaranteed the existence of r controllers for each switch, they did not distinguish master from backup controllers because the master selection was outside their papers scope.

Similar to [7, 8], our present study is based on the assignment of backup NFs to enhance network reliability. However, unlike [7,8], our network functions cannot be both main and backup simultaneously. Moreover, all of the aforementioned studies, related to the placement of gateways, take as a reference the LTE network architecture. In this paper, we propose a more revolutionary approach based on the recent standard of the 3GPP for 5G networks [9]. Furthermore, we analyze the UPFP problem by taking into account parameters such as reliability, latency and relocations. Thus, our paper makes the following **contributions**: 1. It addresses the UPFP problem in the 5G architecture standardized by the 3GPP. 2. It incorporates reliability metrics into the mobile gateways (i.e., UPFs) placement problem. 3. It proposes two MILP to determine the optimal locations of UPFs by considering relocations, latency and reliability metrics. 4. It conceives a strategy that allows for providing resilience against multiple failures while reducing the number of backup UPFs.

3 5G User Plane Reference Architecture

The first standard for the 5G system architecture was defined by the 3GPP in Technical Specifications (TS) 23.501 [9] and 23.502 [10]. This architecture is a (r)evolution of the current 4G network because many of its NFs are the result of the decomposition of some functions executed by nodes of EPC networks, whereas others are entirely new.

The 5G user plane is comprised of UPFs. These NFs can be distributed and deployed closer to the User Equipment (UE) to meet increasing traffic demands while serving low-latency applications hosted at the edge. UPFs are in charge of processing data plane packets between the (Radio) Access Network ((R)AN) and the Data Network (DN). Moreover, they provide access control, packet routing

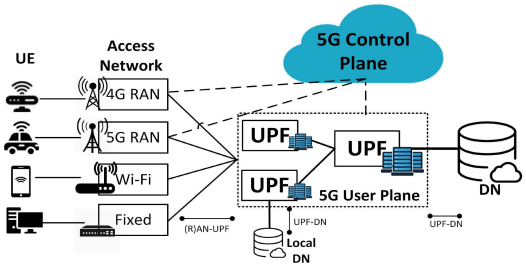


Fig. 1. 5G converged architecture [11].

and forwarding, and Quality of Service (QoS) handling. The UPFs act as anchor points for intra/inter-radio access technology mobility as well as an external Protocol Data Unit (PDU) session point for interconnecting to the DN. To be able to perform these functions, they rely on the Session Management Functions (SMFs) located in the control plane. The SMFs select, manage and control the UPFs to establish PDU sessions. Figure 1 depicts the 5G user plane architecture and its interaction with the access and data networks and control plane.

4 Problem Formulation

Increasing traffic demands along with the stringent requirements of forthcoming services in terms of latency and reliability entail further network transformation. Specifically, low-latency requirements demand the placement of NFs (e.g., UPFs) closer to users at the network edge. Thus, the network response time as well as links congestion can be reduced. In addition, reliability requires the deployment of more NFs to provide higher resilience against failures. This situation implies an increase in the number of UPFs that must be deployed, which translates to increased costs and UPF relocations. Relocations not only degrade QoS but also increase operational costs because of the additional signaling exchanged among NFs to maintain or reestablish PDU sessions. In this context, novel optimization models for the UPFP that comprise latency, reliability and relocation metrics are mandatory for ensuring QoS while reducing deployment and operational costs.

In this section, two optimization models are presented to tackle the UPF placement. The main objective of these models is to determine the optimal placement for virtual instances of UPFs given a set of possible locations; thus, costs are reduced while service requirements of latency and reliability are satisfied. The set of locations may comprise Edge Nodes (ENs) and DCs facilities already deployed by network operators. Moreover, the network model and used notation are also introduced.

4.1 Network Model

The 5G network topology is represented as a graph $G(N; E)$, where N is the set of network nodes and E the links among them. The set of network nodes is

formed by UPF candidate placements (N_c) and access nodes (N_r), which can be fixed and/or radio access technologies. Let L_{rc} denote the shortest distance among access nodes and UPF candidate placements, measured in terms of propagation delay, and L_{req} denote the maximum permissible latency between them. Furthermore, K_u represents the minimum number of backup UPFs to which the access nodes must be assigned to meet reliability requirements. The used notation is summarized in Tables 1 and 2.

Table 1. Sets and parameters

Notation	Description
N_r	Set of access nodes
N_c	Set of UPF candidate placements
d_r	Traffic demand at each access node
C_u	Capacity of each UPF
α	Percentage of the UPF capacity to be occupied
L_{rc}	Latencies between access nodes and UPF candidate placements
L_{req}	Latency requirement between access nodes and UPFs
K_u	Minimum number of backup UPFs to comply with reliability requirements
h_{ij}	Average frequency of handovers between access nodes i and j
F_c	Fixed cost of deploying a UPF at candidate node c
F_h	UPF relocation cost

4.2 Model 1: Cost-aware User Plane Function Placement (CUPFP)

A minimum number of deployed NFs considerably reduce deployment and operational costs. Thus, the main objective of the CUPFP model is to determine the minimum number of UPFs to be deployed while satisfying the service requirements of latency and reliability. Accordingly, the CUPFP problem can be formulated as follows:

$$\text{Min} \sum_{\forall c \in N_c} F_c \cdot (x_c + y_c) \tag{1}$$

s.t.:

$$x_c + y_c \leq 1 \qquad \forall c \in N_c \tag{2}$$

$$p_{rc} \leq x_c \qquad \forall r \in N_r, \forall c \in N_c \tag{3}$$

$$b_{rc} \leq y_c \qquad \forall r \in N_r, \forall c \in N_c \tag{4}$$

$$p_{rc} \geq x_c \qquad \forall r \in N_r, \forall c \in N_c\text{: } Loc_r = Loc_c \tag{5}$$

$$\sum_{\forall c \in N_c} p_{rc} = 1 \qquad \forall r \in N_r \tag{6}$$

Table 2. Binary Variables

Notation	Description
x_c	1 if there is a main UPF installed at node c, $c \in N_c$
y_c	1 if there is a backup UPF installed at node c, $c \in N_c$
z_c	1 if backup UPF at node $c, c \in N_c$, shares its capacity
p_{rc}	1 if access node r, $r \in N_r$, has a main UPF installed at node c, $c \in N_c$
b_{rc}	1 if access node r, $r \in N_r$, has a backup UPF installed at node c, $c \in N_c$
$w_{rcc'}$	1 if access node r, $r \in N_r$, with main UPF at node c, $c \in N_c$, has a backup UPF at node c', $c' \in N_c$
a_{ijc}	1 if access node i or j, $i, j \in N_r$, is assigned to a main UPF installed at node c, $c \in N_c$
k_{ijc}	1 if access node i or j, $i, j \in N_r$, is assigned to a backup UPF installed at node c, $c \in N_c$

$$\sum_{\forall c \in N_c} b_{rc} \geq K_u \qquad \forall r \in N_r \quad (7)$$

$$z_c \leq y_c \qquad \forall c \in N_c \quad (8)$$

$$w_{rcc'} = p_{rc} \wedge b_{rc'} \qquad \forall c, c' \in N_c, \forall r \in N_r \quad (9)$$

$$\text{if } z_c = 1 \Rightarrow \sum_{\forall r \in N_r} d_r \cdot w_{rcc'} \leq C_u / K_u \qquad \forall c, c' \in N_c \quad (10)$$

$$\text{if } z_c = 0 \Leftrightarrow \sum_{\forall c \in N_c} \sum_{\forall r \in N_r} d_r \cdot w_{rcc'} \leq C_u \qquad \forall c' \in N_c \quad (11)$$

$$\sum_{\forall r \in N_r} d_r \cdot p_{rc} \leq \alpha \cdot C_u \qquad \forall c \in N_c \quad (12)$$

$$L_{rc} \cdot (p_{rc} + b_{rc}) \leq L_{req} \qquad \forall r \in N_r, \forall c \in N_c \quad (13)$$

$$x_c, y_c, p_{rc}, b_{rc}, w_{rcc'} \text{ binary} \qquad \forall r \in N_r, \ \forall c \in N_c \quad (14)$$

The objective function, Eq. (1), is aimed at minimizing the deployment cost by taking into account the number of main and backup UPFs to be deployed and their location-dependent cost (F_c). The latter may include other costs, e.g., equipment and operation costs, according to network operator preferences. Equations (2) to (14) define the constraints of the optimization problem.

Inequality (2) ensures that at a specific candidate location just can be placed a main or backup UPF, but not both at the same time. The distinction between main and backup UPFs allows energy saving. As in normal network conditions (no-failure scenarios), the backup UPFs do not have any access nodes assigned; they can be instantiated only when failures occur. In addition, constraints (3) and (4) indicate that an access node cannot be assigned to a candidate location where there is not placed either a main or backup UPF. Moreover, Eq. (5) restricts the assignment of an access node to a specific UPF if this UPF has

been placed at the same location. Specifically, if the access node location has a main UPF, then it must be assigned to it.

Constraint (6) ensures that the access nodes demands are served by exactly one main UPF at a given time. Note that the access nodes could have more than one main UPF assigned if their demands were split by service type or other criteria; however, considering their demands as a whole was preferred to simplify the problem formulation. Additionally, to guarantee the service reliability requirement, constraint (7) was defined. It ensures that the access nodes are assigned to at least the minimum number of backup UPFs (K_u) necessary to provide the required level of reliability. Thus, the user plane can resist against a maximum number of K_u UPF failures by mitigating service interruption.

Because not all UPFs will fail simultaneously, the access nodes that do not belong to the same main UPF could share the capacity of their assigned backup UPF. Therefore, a backup UPF could share its capacity as long as, in the case of K_u failures, its capacity is sufficient to serve the assigned access nodes of the K_u-failed main UPFs. Thus, the number of UPFs for deployment can be reduced by sharing the capacity of the backup UPFs. Equations (8)-(11) express system constraints on sharing backup capacity. Constraint (8) indicates that only the backup UPFs can share their capacity, whereas Eq. (9) expresses the relationship between a main and backup UPF of an access node. Because constraint (9) is nonlinear, it requires further transformation to be linearized. Thus, it can subsequently be replaced with the following expressions: $w_{rcc'} \leq p_{rc}$, $w_{rcc'} \leq b_{rc'}$ and $w_{rcc'} \geq p_{rc} + b_{rc'} - 1$.

Knowing beforehand which exact combination of UPFs will fail at a given time and the capacity occupied in the backup UPFs by their access nodes is almost impossible. To overcome this limitation, the following assumption was made: *if a backup UPF shares its capacity, the total demand of its access nodes that belong to the same main UPF cannot exceed the backup capacity divided by the number of failures to which the system must resist* (see Eq. (10)). Thus, in the case of K_u main UPF failures, the backup UPFs will be able to attempt all the demands of the affected access nodes. By contrast, if a backup does not share its capacity, then its total capacity cannot be exceeded (see Eq. (11)). Note that constraints (10) and (11) are nonlinear and they can be equivalently expressed in a linear form as follows:

$$\sum_{\forall r \in N_r} d_r \cdot w_{rcc'} \leq C_u/K_u + M_1 \cdot (1 - z_c) \qquad \forall c, c' \in N_c \qquad (15)$$

$$\sum_{\forall c \in N_c} \sum_{\forall r \in N_r} d_r \cdot w_{rcc'} \leq C_u + M_2 \cdot z_c \qquad \forall c' \in N_c \qquad (16)$$

$$\sum_{\forall c \in N_c} \sum_{\forall r \in N_r} d_r \cdot w_{rcc'} \leq C_u + \varepsilon + M_3 \cdot (1 - z_c) \qquad \forall c' \in N_c \qquad (17)$$

where M_1, M_2 and M_3 are sufficiently large constants and $\varepsilon > 0$ is a lower bound.

Constraint (12) ensures that the capacity of the main UPFs is not exceeded, where α is the maximum UPF capacity to be occupied by the access nodes to avoid slowing the UPFs performance. Additionally, expression (13) guarantees that an access node is not assigned to either a main or backup UPF if the latency requirement is not satisfied. Finally, Eq. (14) indicates that x_c, y_c, p_{rc}, b_{rc}, and $w_{rcc'}$ are binary variables.

4.3 Model 2: Mobility-aware User Plane Function Placement (MUPFP)

Unlike the CUPFP model, which only considers deployment costs, the MUPFP is aimed at jointly optimizing the deployment and operation costs by considering the effects of user mobility on UPF relocations. Hence, the main objective of the MUPFP is not only to determine the optimal location for the UPFs to minimize the number of UPFs deployed but also the number of UPF relocations.

As previously stated, UPF relocations occur when a user moves between two access nodes that are served by different UPFs. Therefore, the occurrence of relocations in either the main or backup UPFs can be indicated using Eq. (18), where a_{ijc} and b_{ijc} are binary variables that express the relationship between two access nodes and their assignment to a UPF, either main or backup.

$$a_{ijc} = p_{ic} \oplus p_{jc},\ k_{ijc} = b_{ic} \oplus b_{jc} \qquad \forall r \in N_r,\ \forall c \in N_c \tag{18}$$

Thus, the MUPFP problem can be formulated as follows:

$$\text{Min} \sum_{\forall c \in N_c} F_c \cdot (x_c + y_c) + \sum_{\forall c \in N_c} \sum_{\forall i \in N_r} \sum_{\forall j \in N_r} F_h \cdot h_{ij} \cdot (a_{ijc} + k_{ijc}) \tag{19}$$

s.t.: (2) to (14), (18)

In this formulation, the first term of the objective function is associated with the number of deployed UPFs. The second term is related to the cost of UPF reallocations (F_h). The number of UPF relocations is determined by the frequency of handovers (h_{ij}) between access nodes served by different UPFs. Thus, the objective function is aimed at optimizing costs by not only considering the costs caused by the number of deployed UPFs but also the costs associated with the occurrence of UPF relocations. This approach will increase the likelihood of having more access nodes served by the same UPF. Therefore, the number of UPFs and their relocations will be reduced. Note that constraint (18) introduces no linearity to our model and must be replaced with the following inequalities: $a_{ijc} \leq p_{ic}+p_{jc}$, $a_{ijc} \geq p_{ic}-p_{jc}$, $a_{ijc} \geq p_{jc}-p_{ic}$, $a_{ijc} \leq 2-p_{ic}-p_{jc}$, $k_{ijc} \leq b_{ic} + b_{jc}$, $k_{ijc} \geq b_{ic} - b_{jc}$, $k_{ijc} \geq b_{jc} - b_{ic}$ and $k_{ijc} \leq 2 - b_{ic} - b_{jc}$.

The computational complexity of the CUPFP model, in terms of its number of variables and constraints, can be expressed as $O(|N_c|^2 \cdot |N_r|)$ whereas the MUPFP has $O(|N_c|^2 \cdot |N_r| + |N_r|^2 \cdot |N_c|)$ variables and $O(|N_c|^2 \cdot |N_r|)$ constraints. Thus the complexity of both models is asymptotically the same.

5 Performance Evaluation

To assess the performance of the proposed solutions, a test scenario was generated. The scenario represents a 5G network topology deployed in a city of $14km$x$16km$, see Fig. 2. Its access network is composed of 32 nodes (i.e., 22 fixed and 10 radio). The radio access nodes represent centralized Baseband Units (C-BBUs) with a maximum service radius of 3 km. For the placement of UPFs, 13 ENs with a maximum processing capacity of 2.5 Tb/s were considered as candidate locations. To evaluate the performance of our solutions, two services from the URLLC category were selected, i.e., mIoT and vehicle-to-infrastructure cooperative sensing. Their demands were generated using the information provided in Table 3 and considering one active PDU session per user; specifically, a total demand of 2.67 Tb/s in the (R)AN was considered.

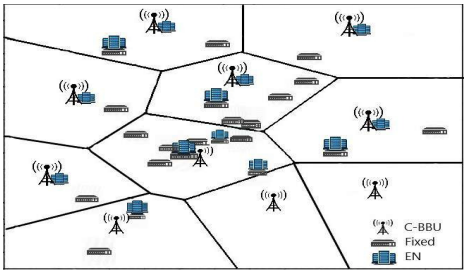


Fig. 2. 5G access network topology.

Table 3. Use cases requirements [11, 12]

Service	**Latency**	**Data Rate per user**	**Density**	**Reliability**
mIoT	$\leq$ 1 ms	$\leq$ 1 Mbps	10^4 $users/km^2$	99.999 %
Cooperative Sensing	$\leq$ 1 ms	$\leq$ 5 Mbps	$\leq$ 100 $users/km^2$	99.999 %

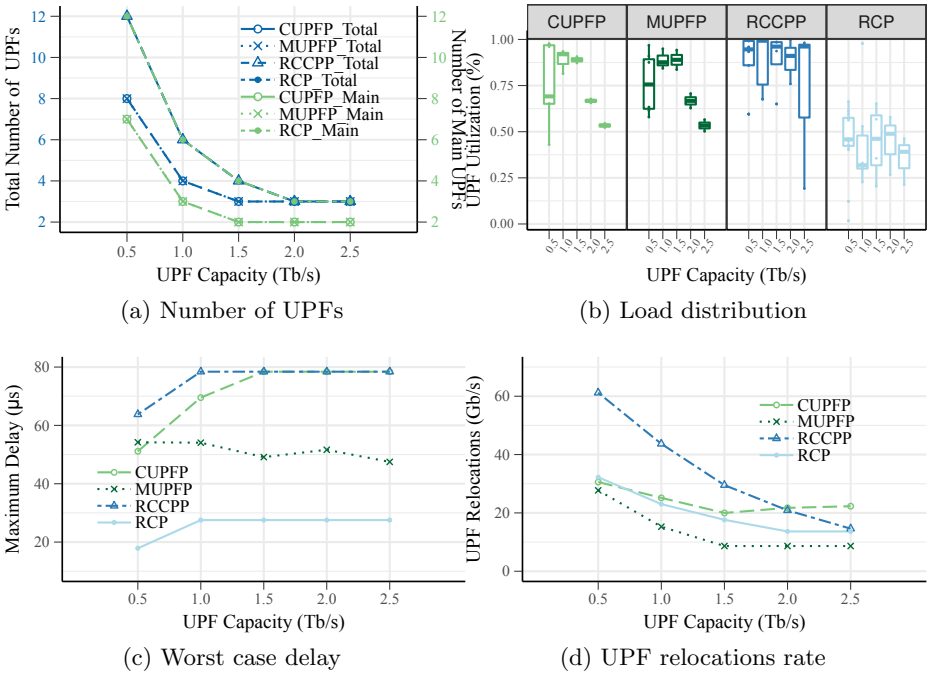
To compare the performance of our solutions with selected relevant studies, the RCP [7] and the unextended version of the Resilient Capacitated Controller Placement Problem (RCCPP) [8] were used as references. We selected the RCP and RCCPP because of their similarities to our models. Their main purpose are to minimize the number of controllers subject to latency, resilience and capacity constraints. To apply these models to the UPFP problem, the controllers were considered UPFs and the switches as access nodes. Additionally, for the RCP implementation a UPF failure probability of 10^{-4} was assumed whereas, in the RCCPP, the intercontroller latency constraint was relaxed. For the implementation of the models, the Python-based package Pyomo was selected along with Gurobi as its underlying solver.

5.1 Analysis of the Results

All the models (i.e., CUPFP, MUPFP, RCCPP and RCP) were evaluated for different values of UPF capacity by considering one level of backup (K_u=1). Their optimal solutions were determined with zero optimality gap and analyzed in terms of the number of UPFs necessary to cover the services demand, load distributions, worst case delays and UPF relocations (see Fig 3).

Number of required UPFs: The total number of UPFs required by all the models is shown in Figure 3a. Additionally, the number of main UPFs obtained by the proposed models and RCP is also represented. At first glance, in all the solutions, it can be observed that the number of required UPFs decreases as the capacity increases. Moreover, the total number of UPFs of the proposed models was always lower than or equal to that of the reference models. This difference is more notable for small values of capacity where the number of UPFs is higher.

For all the values of capacity, the CUPFP and MUPFP models always obtained similar results, either in terms of total or main UPFs. Moreover, their numbers of main UPFs were always lower than the total. Therefore, the proposed solutions are more cost-effective in terms of the numbers of UPFs and resources consumption than the RCCPP and RCP models. Thus, the total numbers of UPFs can be considerably



(a) Number of UPFs (b) Load distribution

(c) Worst case delay (d) UPF relocations rate

Fig. 3. Performance comparison against UPF capacity variation

reduced by sharing the backups capacity. Specifically, by placing one backup UPF the reliability requirement of all access nodes was satisfied. Moreover, the distinction between main and backup UPFs allows energy saving because the backup UPFs can be instantiated only when failures occur.

Load distribution: For our solutions and the RCP, the load distribution was measured only in the main UPFs, whereas in the RCCPP, all the UPFs were included. In Fig. 3b, our proposed solutions can clearly be observed to outperform the reference models for all values of capacity analyzed, with the exception of C_u=0.5 Tb/s. Their maximum imbalance obtained was always below 20 % except for C_u=0.5 Tb/s where the imbalance in the CUPFP was around 60 %. By contrast, the RCCPP and RCP lowest imbalance was always above 25 %, and for C_u=2.5 Tb/s, the imbalance nearly reached 100 % in the RCCPP.

Moreover, our models provided a UPF average utilization between 50 and 90 % whereas this metric was always above 90 % for the RCCPP and below 50% for the RCP what can lead to overload and underutilized UPFs, respectively. These satisfactory outcomes are because of the utilization of the α factor in Eq. (12), which restricts the capacity to be occupied in the main UPFs, thereby allowing for an enhanced distribution of load. This factor was determined in function of the total demand in the (R)AN and the expected number of UPFs for each value of capacity. In addition, the load distributions obtained with the CUPFP and MUPFP models were quite even, although CUPFP outperforms MUPFP for capacities values higher than 1 Tb/s.

Maximum delay: The maximum propagation delay between UPFs and access nodes was calculated in terms of the Euclidean distance divided by the speed of light $2x10^8$ m/s, assuming optical fiber as the underlying transport. To meet the latency requirement of 1 ms, the overall latency (Round-Trip-Time (RTT)) in the (R)AN, should not exceed 0.5 ms [13]. Therefore, the propagation and processing delays in the segment (R)AN-DN cannot excess 0.5 ms. Considering that SGWs and PGWs have a processing time of 100 μs [14], a total processing time of 300 μs in UPFs and DNs is assumed. Moreover, the propagation latency between UPFs and local DNs is negligible because they are assumed to be collocated. Taking the previous analysis into account, an RTT of 200 μs for the propagation latency between UPFs and access nodes (L_{req}) was considered.

Figure 3c represents the worst-case propagation latency between access nodes and UPFs, in one way. The best performance was provided by the RCP, with maximum delays below 30 μs. This is because the RCP is aimed at not only minimizing the number of UPFs but also the routing cost. In addition, the CUPFP and RCCPP obtained similar results with maximum delays up to 78 μs. Notably, a substantial difference did not exist between the RCCPP and RCP and the proposed solutions, despite the number of active UPFs being higher in the reference models. Furthermore, in all the models, the worst-case delay was always below the established threshold ($\leq$ 100 μs in one way).

UPF relocations: The rate of UPF relocations was determined by considering the services data rate and a maximum frequency of handovers between BBUs of [350, 550] HO/s, according to user density. For user mobility, a simple model in which users move with constant speed and direction, in an area, was assumed. In Fig. 3d, the MUPFP can be observed to be the optimal solution because its objective function is aimed at optimizing the occurrence of relocations; by contrast, the RCCPP provides the worst results. In all solutions, the rate of UPF relocations decreases as the number of UPFs reduces. Additionally, a comparison between CUPFP and MUPFP models revealed a remarkable difference in terms of relocations, despite the models having the same number of active UPFs. This result demonstrates the importance of considering user mobility patterns during the placement. Furthermore, this consideration guarantees enhanced QoE without incurring additional costs, measured in terms of the number of deployed UPFs.

6 Conclusion

In this paper, we proposed two MILP models to address the placement of UPFs in 5G networks. The proposed solutions are aimed at not only minimizing the number of UPFs but also their relocations while satisfying the service requirements of latency and reliability. The obtained results showcase the effectiveness of the proposed approaches. Specifically, the number of UPFs to be deployed can be considerably reduced by sharing the capacity of backup UPFs, and UPF relocations can be diminished by considering user mobility and differentiating the main UPFs from the backups.

In future works, we will consider the design of heuristic solutions for the UPFP as well as their evaluation in different settings. Additionally, dynamic optimization of nodes assignment and UPFs placement to adapt to variations in traffic and user locations will be addressed. Furthermore, we intend to solve the placement problem of 5G UPFs by considering the existence of several network slices to optimize resource utilization when services with different requirements coexist.

Acknowledgment

This work has been supported by the Ministerio de Economía y Competitividad of the Spanish Government under the project TEC2016-76795-C6-1-R and through a predoctoral FPI scholarship.

References

1. Huawei Technologies Co.: 5G Network Architecture A High-Level Perspective (2016), https://www.huawei.com/minisite/hwmbbf16/insights/5G-Nework-Architecture-Whitepaper-en.pdf
2. Taleb, T., Ksentini, A.: Gateway relocation avoidance-aware network function placement in carrier cloud. In: Proceedings of the 16th ACM international conference on Modeling, analysis & simulation of wireless and mobile systems. pp. 341–346. ACM (2013)
3. Ksentini, A., et al.: On using SDN in 5G: the controller placement problem. In: Global Communications Conference (GLOBECOM). pp. 1–6. IEEE (2016)
4. Bagaa, M., Taleb, T., Ksentini, A.: Service-aware network function placement for efficient traffic handling in carrier cloud. In: 2014 IEEE Wireless Communications and Networking Conference (WCNC). pp. 2402–2407 (Apr 2014)
5. Taleb, T., Bagaa, M., Ksentini, A.: User mobility-aware virtual network function placement for virtual 5G network infrastructure. In: 2015 IEEE International Conference on Communications (ICC). pp. 3879–3884. IEEE (Jun 2015)
6. Liu, J., Shi, Y., Zhao, L., Cao, Y., Sun, W., Kato, N.: Joint placement of controllers and gateways in SDN-Enabled 5G-Satellite Integrated Network. IEEE Journal on Selected Areas in Communications **36**(2), 221–232 (2018)
7. Tanha, M., Sajjadi, D., Pan, J.: Enduring Node Failures through Resilient Controller Placement for Software Defined Networks. In: Global Communications Conference (GLOBECOM), 2016 IEEE. pp. 1–7. IEEE (2016)
8. Tanha, M., Sajjadi, D., Ruby, R., Pan, J.: Capacity-aware and Delay-guaranteed Resilient Controller Placement for Software-Defined WANs. IEEE Transactions on Network and Service Management (2018)
9. 3GPP: TS 23.501- System Architecture for the 5G System; Stage 2, http://www.3gpp.org/ftp/Specs/archive/23_series/23.501/23501-f00.zip
10. 3GPP: TS 23.502- Procedures for the 5G System; Stage 2, http://www.3gpp.org/ftp/Specs/archive/23_series/23.502/23502-f10.zip
11. 5G Americas: 5G Network Transformation. Tech. rep., 5G Americas (2017), http://www.5gamericas.org/files/3815/1310/3919/5G_Network_Transformation_Final.pdf
12. NGMN Alliance: Perspectives on Vertical Industries and Implications for 5G. Tech. rep., NGMN Alliance (2016), https://www.ngmn.org/fileadmin/user_upload/160922_NGMN_-_Perspectives_on_Vertical_Industries_and_Implications_for_5G_final.pdf
13. Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A.I., Dai, H.: A survey on low latency towards 5G: RAN, core network and caching solutions. IEEE Communications Surveys & Tutorials **20**(4), 3098–3130 (2018)
14. Tawbeh, A., Safa, H., Dhaini, A.R.: A hybrid SDN/NFV architecture for future LTE networks. In: 2017 IEEE International Conference on Communications (ICC). pp. 1–6. IEEE (2017)