Research Article

A Robust Optimization Based Energy-Aware Virtual Network Function Placement Proposal for Small Cell 5G Networks with Mobile Edge Computing Capabilities

Bego Blanco, Ianire Taboada, Jose Oscar Fajardo, and Fidel Liberal

Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU), Alameda Urquijo s/n, 48013 Bilbao, Spain

Correspondence should be addressed to Bego Blanco; begona.blanco@ehu.eus

In the context of cloud-enabled 5G radio access networks with network function virtualization capabilities, we focus on the virtual network function placement problem for a multitenant cluster of small cells that provide mobile edge computing services. Under an emerging distributed network architecture and hardware infrastructure, we employ cloud-enabled small cells that integrate microservers for virtualization execution, equipped with additional hardware appliances. We develop an energy-aware placement solution using a robust optimization approach based on service demand uncertainty in order to minimize the power consumption in the system constrained by network service latency requirements and infrastructure terms. Then, we discuss the results of the proposed placement mechanism in 5G scenarios that combine several service flavours and robust protection values. Once the impact of the service flavour and robust protection on the global power consumption of the system is analyzed, numerical results indicate that our proposal succeeds in efficiently placing the virtual network functions that compose the network services in the available hardware infrastructure while fulfilling service constraints.

1. Introduction

In this new era of an always-connected and hypercommunicated society, 5G targets a set of ambitious requirements in terms of network performance-related key performance indicators such as very low delays and high data rates [1–3]. For this reason, from an operator's perspective, finding a tradeoff between user performance improvement and CAPEX/OPEX reduction becomes essential.

In this context, Cloud-Enabled Radio Access Networks (CE-RANs) become a feasible solution to enhance user experience and decrease total costs [4]. CE-RAN architecture leverages network function virtualization (NFV) [5–7] to split the radio protocol stack between physical and virtualized functions, the latter running in edge clouds. This edge cloud may also be used to place service instances at the edge of the mobile network and in the proximity of end users. Thus, this inclusion of edge cloud service capabilities in the network not only improves quality of service but also simplifies deployment and management in multitenant scenarios [8].

Through the deployment of Virtualized Network Functions (VNFs) on Virtual Machines (VMs), operators will execute edge radio and service VNFs inside data-centers on commodity hardware with the advantage of reducing capital expenses. In this novel paradigm, general-purpose computing in networks has been realized along with the virtualization of network functions, which enables the automation of network service provisioning and management. In this work we focus on an emerging 5G scenario, a multitenant cluster of small cells (SCs) that provides edge radio and service functions by means of virtualization technologies. In particular, we employ cloud-enabled SCs that integrate microservers for virtualization execution, equipped with IT resources (CPU, RAM, and storage) and additional hardware appliances such as GPUs, DSPs, or FPGAs (Figure 1). Thus, in this whole picture, we could define a network service (NS) as a collection of VNFs required to deploy a complete 5G mobile service (e.g., innovative video content delivery at the edge integrated with radio monitoring and management) for the end users of the SC network operator.

SC: small cell   PNF: physical NF
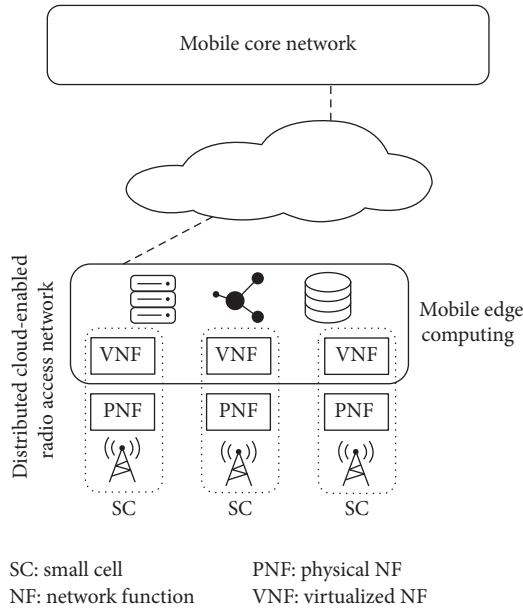NF: network function   VNF: virtualized NF

Figure 1: Cloud-enabled RAN architecture based on small cells.

Nevertheless, the success of the presented novel approach undoubtedly depends on the VNF placement strategy used to allocate VNFs into the available resources given a set of service constraints. However, defining efficient mechanisms to allocate the software-based components of NS onto the networked data-centers (set of microservers that conform the cloud-enabled SC cluster in our case) is a challenging task. Therefore, the relevance of designing intelligent VNF placement algorithms becomes fundamental, while it is not defined in standards [9].

Besides, current trends towards green information and communication technologies target the improvement of energy consumption in all the nodes including the network nodes themselves. In this paper, we will particularly discuss how to derive an energy-aware VNF placement solution for a cluster of cloud-enabled SCs in a 5G deployment scenario. With this purpose, we will deal with the objective of minimizing power consumption of both the microservers and SDN-enabled switches to deploy the overall virtualized infrastructure needed (including virtual switches within each SC and physical ones interconnecting SCs).

In order to address the VNF placement problem under the aforementioned novel 5G scenario, this paper proposes a placement mechanism that minimizes power consumption subject to bit rate and latency constraints imposed by the network services and is aware of the available underlying resources. Existing literature provides a wide variety of VNF placement approaches, ranging from accurate Integer Linear Programming (ILP) formulations to heuristic algorithms. However, these solutions are more focused on the use of NFV over big centralized data-centers rather than over small edge clouds devoted to provide radio access and edge networking services. In addition, the analyzed studies rarely consider uncertainty in the definition of the problem.

Hence, the main contribution of this work is the design of an energy-aware VNF placement expert system for

a CE-RAN environment with 5G SCs. We use constraint programming to solve the discrete optimization problem of resource assignment, combined with robust optimization techniques (RO) [10], to develop an energy-aware VNF placement policy. Thus, our approach differentiates from others in the use of RO, which allows us to generate a suboptimal placement policy that enables the introduction of uncertainty in the problem. In particular, we introduce service demand uncertainty in order to capture the variable nature of traffic flows size and per-service job execution burden.

The paper has been organized in six sections. Section 2 reviews the related work. In Section 3, we describe the VNF placement problem under a 5G environment. Section 4 presents our placement proposal using the RO approach, previously providing the problem model. Next, in Section 5, we analyze the performance of the proposed placement solution. Finally, Section 6 gathers the conclusions of this work.

2. Related Work

Nowadays, mobile operators see network virtualization as the key towards enabling flexible 5G network architecture. For this reason, the problem of VNF placement has been attracting increasing attention during the last years [11–25]. Nonetheless, the optimal placement of virtual elements to physical resources aimed at minimizing power consumption with service performance constraints for the 5G scenario under study is a complex issue.

VNF placement is a resource allocation discrete or combinatorial optimization problem that can be considered equivalent to a knapsack problem. In consequence, it can be solved using exact methods or heuristics depending on the number of VNF instances and the complexity of the resources in competition. As typically provided in the literature (see [12–14]), an ILP approach optimally solves the placement problem in reasonable time when the number of service instances is low. In reference to heuristics-based placement proposals, the variety of tools employed for solving the problem is huge; for example, [15] uses a greedy algorithm whereas [16] uses a binary search heuristic.

Even though some of the aforementioned relevant works deal with energy-related optimization taking into account delay bounds (e.g., the objective in [12, 13] is minimizing the number of CPUs constrained by latency), those solutions do not usually consider optimal/suboptimal power consumption itself. However, it is worth mentioning the few works we have found deal with power-based VNF placement proposals [17–21] for our case study. Reference [17] aims at minimizing power consumption allowing users to meet delay requirements using a genetic algorithm approach, while [18, 19] propose heuristic algorithms that give power reduction improvement. Besides, [20] presents a suboptimal placement solution that combines traffic and energy cost optimization derived by means of a Markov approximation with matching theory. Apart from that, research works in [22–24] go a step forward and cope with the migration problem of active microservers needed in low traffic conditions, to turn off or suspend microservers so as to achieve energy savings.
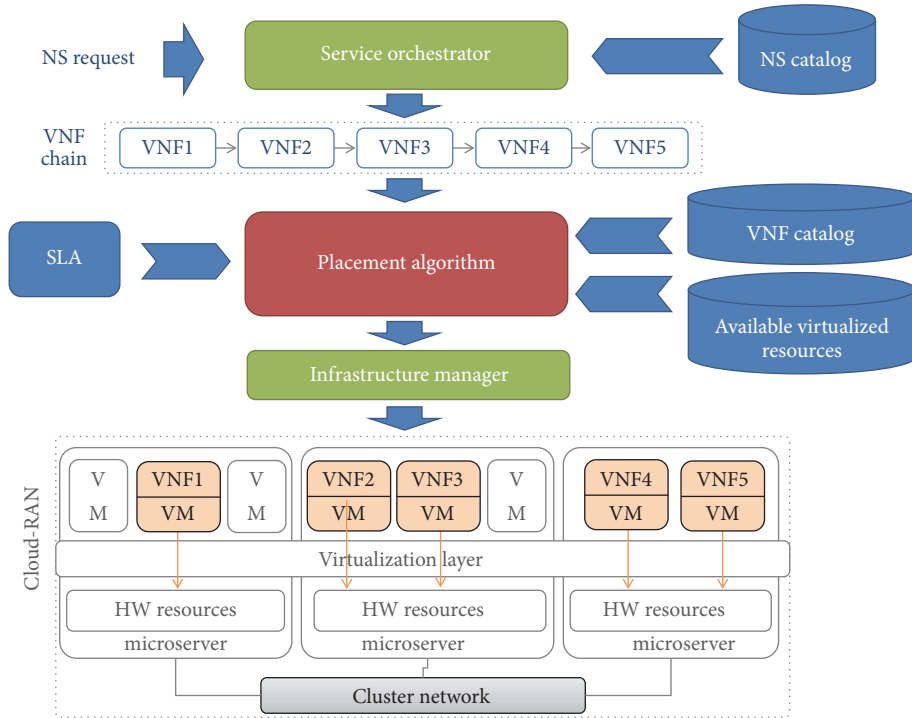
FIGURE 2: VNF placement process.

Moreover, a common assumption of the different optimization models proposed to solve different versions of the VNFs placement problem is that input data is perfectly known, which is difficult in practice. Unfortunately, small deviations in input data values may usually lead to situations where a previously found optimal solution is not even feasible anymore, and consequently the presence of uncertain data may produce useless placement solutions. In order to cope with uncertain input parameters, RO [10] is applied to solve optimization problems. Indeed, RO has been successfully employed in different contexts, such as in resource allocation in OFDM networks [26] with uncertainty in channel state information, in cloud resource provisioning without perfect knowledge of demands and price [21], and even in the virtual network embedding problem on a physical network substrate in recent works [27–30] that usually deal with uncertain service demands.

Nevertheless, to the best of our knowledge, except the work presented in [21], there is no VNF placement proposal based on RO. Therefore, to achieve our goal of designing an energy-aware placement solution using RO approach that considers service demand uncertainty, [21] has been relevant. Authors develop a novel power-based placement solution with uncertainty in VNFs CPU demands based on RO theory. To this aim, they take into account the physical servers and network resources along with their energy efficiency. Nonetheless, they focus on the core network part whereas our research is oriented to combined deployment of radio/service network functions at the edge. Additionally, the placement problem considered in this paper takes into account a heterogeneous edge cloud made up of a cluster of interconnected SCs.

3. Problem Description

The networking environment targeted in our work is a 5G distributed edge cloud. This environment uses SCs that evolve to build multitenant CE-RANs. The enhanced SCs deployment is offered to multiple operators/service providers to serve their network services in the virtualized infrastructure and engage them in a multitenant system. Figure 2 shows the NS request and subsequent placement decision process in this joint radio-cloud architecture.

In such an architecture, whenever a tenant needs to deploy a new service, it makes a request to the Service Orchestrator functional element, which is implemented as the NFV Orchestrator (NFVO) in ETSI-MANO terminology. The orchestrator analyzes its NS catalog to extract the correspondent service chain. A service chain is the sequence of VNFs that composes the requested 5G service.

The VNF chain extracted from the catalog is later used as an input for the placement algorithm together with the service constraints imposed by the key performance indicators (KPI) of the corresponding Service Level Agreement (SLA) between the tenant and the CE-RAN operator. In our work, the SLA KPIs include the following network parameters: maximum accepted latency for the NS, aggregated user bit rate, and a robust protection parameter. This protection parameter is related to the uncertainty of the user traffic demand. We assume that the network service has a mean aggregated traffic demand that is used to size the necessary resources in order to serve tenant's user. But, along the service time, the user traffic demand fluctuates around this mean value, affecting the amount of resources needed. The robust protection parameter is therefore related to the expected peak

traffic demand deviation and will be used to overestimate the allocation of virtualized resources, in order to be prepared for eventual user peak demands.

Next, the placement algorithm inspects the VNF catalog for the parametrization of the VNF instances needed to match the SLA and checks the available resources provided through the virtualized infrastructure. VNFs are characterized by the use of resources (i.e., number of cores, RAM, and storage) according to the aggregated bit rate (i.e., characterized aggregated traffic flows of a NS) served by the NS they belong to. Moreover, the service latency of each VNF also depends on the traffic load supported by the VNF instance. Lastly, virtualized resources may also include other hardware appliances, such as hardware accelerators (HWA), that can improve the performance of a VNF in terms of latency. This way, these additional appliances help matching SLA with the tenant, but at a higher energy cost.

All the VNFs composing the service chains of the requested network services must be allocated in the available virtualized resources, including a given number of CPUs, hardware appliances, RAM and storage space, and network bandwidth. We assume that each core runs a single VM that is devoted to the execution of a single VNF instance, but one VNF instance can scale from one to many VMs upon the service requirements of the NS. Besides, the transmission latency along the network topology between two VNFs allocated in different microservers depends on the traffic load of the involved links and the size of the flow to be transmitted. With this information, the placement algorithm assigns each VNF instance to one or more VMs with the objective of minimizing the energy consumption of the whole system while matching the latency constraints for the NS.

Finally, the Infrastructure Manager functional element, Virtualized Infrastructure Manager (VIM) in ETSI-MANO terminology, receives the outcome of the placement decision and maps each VNF to the assigned resources.

It is important to point out that a VNF chain includes radio-related and service-related instances. The placement algorithm introduced in this work considers a scenario with several tenants, each of them requesting one or more NSs, which share the available resources. Considering the aforementioned communication particularities of 5G, Figure 3 shows an example of a simple complete placement result of a multitenant environment.

This example considers a distributed edge cloud shared by two tenants. Tenant 1 offers two network services, the first one chaining a virtualized firewall (vFW), an intrusion detection function (vID), a watermarking process (vWM), and finally a caching service (vCache). The second NS of Tenant 1 chains the vFW and a video analytics (VA) functionality for online surveillance. Tenant 2 offers a single network service than chains the vFW, a vCache service, and a virtual transcoding unit (vTU). All NSs are associated with their correspondent latency constraint.

In the example, both NSs belonging to the same tenant share the radio VNF (small cell VNF or SCVNF). In this way, the common radio operations are performed in a coherent way between the different edge services of a tenant, and the data paths of the specific service instances per tenant are split.
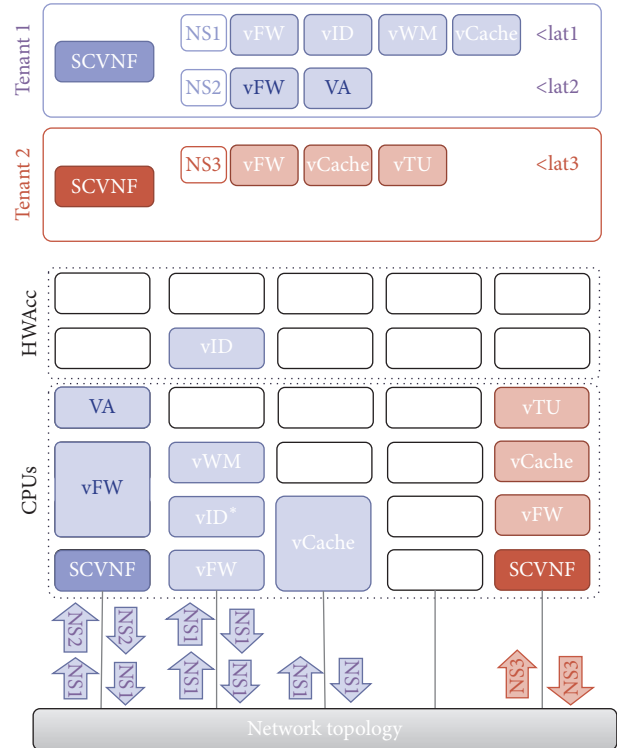

FIGURE 3: Multitenant placement example. ∗ indicates that "vID" uses a hardware acceleration appliance.

The SCVNF of each tenant will be shared by all the NSs of this tenant, which means that the use of resources of the SCVNF must then consider the sum of the aggregated flows of the NSs of the same tenant.

The distinctive features of 5G communications have two important implications over the placement algorithm. Firstly, every tenant allocates a unique SCVNF instance at the edge cloud in order to support the control plane and data plane operations of all the SCs associated with that tenant. And, secondly, since the SCVNF manages the uplink and downlink traffic between the mobile users and the core network, it leads to circular forwarding paths when including edge service instances. In this paper, we consider that the core network functionalities of the mobile network operator are outside the CE-RAN, which means that the CE-RAN only deploys radio and edge service instances. In order to keep the data path between the mobile users and the core network functions, all the NSs need to start and end at the SCVNF, which will handle the user plane traffic to/from the core network.

The example also shows how some VNFs need more than one VM to execute the associated functionalities and how one VNF uses a HWA to meet the latency constraint of the server.

Section 5 will show the result of more complex and complete placement experiments. For the sake of generality, this paper does not focus on the specific modelling of the traffic patterns, the energy consumption schemes, or the VNF network service performance. Section 5 provides some reference outcomes on the potential benefits of the introduced VNF placement algorithm in the considered network scenario, but

the applicability of the solution to other scenarios should be fed by experimental modelling data specific to the target scenario.

4. Robust Constraint-Based VNF Placement Proposal

Once the 5G VNF placement problem under study is presented, in this section we propose a placement solution to that problem using the RO approach. To this end, first, we formulate the energy-aware VNF placement model in the next subsection. Then, we provide our RO-based placement proposal in Section 4.2.

4.1. Problem Modelling. As stated before, the VNF placement problem addressed in this work focuses on the assignment of the VNFs composing each NS to the provided 5G network infrastructure. Therefore, it constitutes a discrete or combinatorial optimization problem, where the minimization of energy consumption constitutes the optimality criterion. Summary of Model Indexes, Sets, Parameters, and Decision Variables reports and summarizes the mathematical notations used for the model indexes, sets, input parameters and decision variables of VNF placement problem that is defined in the following.

The Problem Model. The energy-aware VNF placement optimization problem:

$$\min_{\pi \in \Pi} \quad P^\pi \tag{1}$$

$$\text{s.t.} \quad d_i^\pi < D_{\max_i} \quad \forall i \in \mathcal{I}, \ \pi \in \Pi \tag{2}$$

$$m_{rx}^\pi < M_{rx} \quad \forall r \in \mathcal{R}, \ x \in \mathcal{X}, \ \pi \in \Pi \tag{3}$$

$$tl_x^\pi < \text{BW}_x \quad \forall x \in \mathcal{X}, \ \pi \in \Pi \tag{4}$$

$$P^\pi = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} P_{ij}^\pi \tag{5}$$

$$P_{ij}^\pi = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} P\text{cpu}_{ij}^{xy} A_{ij}^{xy} + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} P\text{switch}_{ij} L_{ij} + \sum_{x \in \mathcal{X}} P\text{sc}_x A_{ij}^{xy} \tag{6}$$

$$d_i^\pi = \sum_{j \in \mathcal{J}_i} d_{ij}^\pi \tag{7}$$

$$d_{ij}^\pi = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} dp_{ij}^{xy} A_{ij}^{xy} + \sum_{x_1 \in \mathcal{X}} \sum_{y_1 \in \mathcal{Y}_{x_1}} \sum_{x_2 \in \mathcal{X}} \sum_{y_2 \in \mathcal{Y}_{x_2}} dl_{ij}^{x_1 y_1, x_2 y_2} A_{ij}^{x_1 y_1} A_{ij+1}^{x_2 y_2} \tag{8}$$

$$\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} A_{ij}^{xy} = 1 \quad \forall x \in \mathcal{X}, \ y \in \mathcal{Y}_x \tag{9}$$

$$m_{rx}^\pi = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} \sum_{y \in \mathcal{Y}} m_{ij}^r A_{ij}^{xy} \tag{10}$$

$$tl_x^\pi = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} \sum_{y \in \mathcal{Y}} T_i A_{ij}^{xy} A_{ij+1}^{x'y}. \tag{11}$$

We consider a set $\mathcal{X}$ of small cells that compose the CE-RAN and a set $\mathcal{Y}_x$ of cores included in the microserver of the SC x. Additionally, we also consider a set of $\mathcal{I}$ network services offered by the CE-RAN operator to the different tenants and a set $\mathcal{J}_i$ of VNFs that form the functional chain of the NS i. Finally, $\mathcal{R}$ represents the set of different resources available in the microservers for the execution of the VNFs (storage, RAM, hardware accelerators, etc.). Hence, our constraint-based expert system must determine in which VM y of a SC x each VNF j of a NS i is located.

Let Π denote the set of all possible power-aware and latency- and resource-constrained VNF placement policies, which decide the mapping of each VNF of available NSs to available VMs in the provided SCs. We formulate our VNF placement optimization problem aimed at minimizing power consumption with delay requirements as (1).

In reference to the energy part, we define the total power consumption P^π as the sum of the energy demand of all the allocated VNFs P_{ij}^π (5). Then, for each VNF instance, the power consumption is expressed in (6) as the sum of the power consumption of VNF j of NS i due to the CPU use of VM y of SC x, the power consumption of the physical switch due to the forwarded traffic of the VNF, and the power consumption of SC x because of being switched on.

The power consumption of the cores that run the VMs depends of the aggregated user traffic served by the correspondent VNF, and this relationship may be nonlinear. It must be also understood that when a VNF instance scales to a higher number of cores, the individual load of the involved CPUs decreases, leading to a lower power consumption per core. Additionally, the energy consumption of the networking devices (e.g., an Ethernet switch) depends on the aggregated flow that traverses the device to forward data packets between SCs and the number of active ports. Finally, it is realized that the power wasted because the SC is active is constant, not affecting the optimization results. Besides, we express per-NS delay in (7), as the sum of per-VNF delays, and the per-VNF delay in (8) as the sum of the processing delays introduced by the execution of the VNF instance, plus the path delay introduced by the traffic going from one VNF to the next one in the chain defined in the correspondent NS.

On the one hand, the processing delay in the CPU grows with the supported user traffic load. On the other hand, we consider that the path/link delay is composed of several link delays from/to VMs to/from the switches plus the delay in the switches. Hence, depending on the adopted placement policy π, each NS will follow a different path through the network. In such a way, if the VNFs belonging to a NS are placed in the same SC the service chain traverses virtual/internal links and the virtual switch of the selected SC, whereas if the VNFs of

a NS are located in different SCs external links and the physical switch is also used, therefore, generally, $dl_{ij}^{x_1 y_1, x_2 y_2} = d_{x_1 y_1, vs} + d_{x_1}^{vs} + d_{vs_{x_1}, s} + d_s + d_{s, vs_{x_2}} + d_{x_2}^{vs} + d_{vs, x_2 y_2}$; but for a given VNF if the next VNF in its chain holds in the same SC, its link delay is only affected because of passing through the virtual switch. Note that in case a VNF is the last VNF in the chain the link delay is null, with the sums referring to the link delay for that j being null.

In order to guarantee that each VM holds only a single VNF, constraint (9) must be satisfied. Nevertheless, we consider that each VNF of a NS can be instantiated as a VM that is executed over several cores of the same SC. Apart from that, as previously detailed in Section 3, the first VNF and the last VNF in a service chain are the same (the SCVNF) in order to perform the specific functions that allow the split between control and data plane required in 5G communications. Therefore, it implies that $A_{i1}^{xy} = A_{ij}^{xJ_i}$.

Finally, the model must also include the constraints referred to as the available resources. In this sense, (3) states that the consumption of the resource r due to the placement of VNFs in the microserver of the SC x, m_{rx}^{π}, must not exceed the total amount of resource M_{rx}, where m_{rx}^{π} is defined in (10). Similarly, (4) constrains the traffic through the network links tl_x^{π} to the maximum capacity of the links BW_x, as the sum of the traffic flows of the network services that must be forwarded from a VNF in one SC to the next VNF in another SC.

4.2. Robust Constraint-Based Solution. As described in Section 2, many existing placement works assume perfect knowledge on input parameters pointing, thus, to the formulation of a deterministic optimization problem. Unfortunately, in the real world, the estimated values of the input parameters may differ due to biased data, unrealistic assumptions, or numerical errors, affecting the obtained optimal solution and its performance. This potential deviation of the nominal or expected input may lead to the violation of the problem constraints and therefore make the obtained solution suboptimal or even unfeasible.

The uncertainty of modelling parameters has been traditionally handled through stochastic programming and sensitivity analysis, but robust optimization techniques have recently arisen as a powerful tool to manage the impact of uncertain input sets. RO seeks an uncertainty-immunized solution with an acceptable performance under the realization of the uncertain inputs, becoming a conservative, worst-case oriented methodology. The main tools of RO are *uncertainty sets* and a *robust counterpart problem* [31]. The uncertainty of the input data is described later in this section, while the robust counterpart problem is the deterministic model previously detailed in Section 4.1.

Here we propose a robust constraint-based placement solution that deals with service demand uncertainty. With this RO approach, we consider a tradeoff between the problem optimization and the protection from deviations caused by input parameters uncertainty. This uncertainty in service demand has an impact on several components on the formulation previously presented in Section 4.1. On the one

hand, we consider that a VM which is running a specific VNF requires an expected use of CPU and thus a certain constant power consumption due to CPU. Nevertheless, the uncertainty in traffic demand may cause uncertainty in the CPU requirement/use and, consequently, in its energy consumption and processing delays. On the other hand, both energy consumption and delays in switches and links depend on that demand uncertainty as well.

In order to obtain a proposal for the placement problem with uncertain service demand, we employ the Soyster [10] method used in the framework of RO. That technique is also known as the worst-case approach, where the solutions are achieved using the most extreme expected values of the uncertain variables. Therefore, that kind of resolution guarantees feasible solutions for any value of the uncertain service demand.

In reference to the modelling of the uncertain traffic demand per NS i, we consider an expected traffic demand, $E[T_i] = T_i$, plus a term due to uncertainty in the service demand, UT_{max_i}. We assume that the random variable UT_{max_i} is symmetrically distributed within $[-\Delta T_i, +\Delta T_i] \cdot T_i$ and with mean zero. Furthermore, the sum of deviations of the uncertain service demand should not exceed the worst-case maximum value T_{max_i}, fulfilling (12). In this way, the value of T_{max_i} controls the tradeoff between the robustness and the impact on the optimization; higher T_{max_i} leads to worse objective function values but protects from more parameter deviations.

$$\frac{UT_i}{\Delta T_i} \leq T_{max_i} \quad \forall i. \tag{12}$$

5. Analysis of the Results

This section presents the results obtained with the simulation of the proposed robust constraint-based expert system for the placement of VNF chains in an emerging 5G distributed edge cloud. As previously introduced in Section 2, the mathematical VNF placement problem model presented in Section 4 is a combinatorial or discrete optimization problem equivalent to the well-known knapsack problem. This means that the decision problem is NP-complete, while the optimization problem is NP-hard [11]. The proposed optimization model has been solved using constraint programming as a powerful paradigm for solving combinatorial search problems [32].

We used Minizinc [33], a free and open-source constraint modelling language, to model the optimization and constraint satisfaction VNF placement problem in a high-level, solver-independent way. Then, Minizinc compiles this model into FlatZinc, a solver input language that is understood by a wide range of solvers. In our case, Gecode solver [34] provided the best performance in the search of the optimal solution.

The experiments that are described and discussed below have been conducted in a single computer with a 2,7 GHz Intel Core i5 processor and 16 GB RAM. This simple configuration provides solutions within minutes for simple experiments with loose constraint satisfaction. When the experiments demand tighter latency or resource constraints,
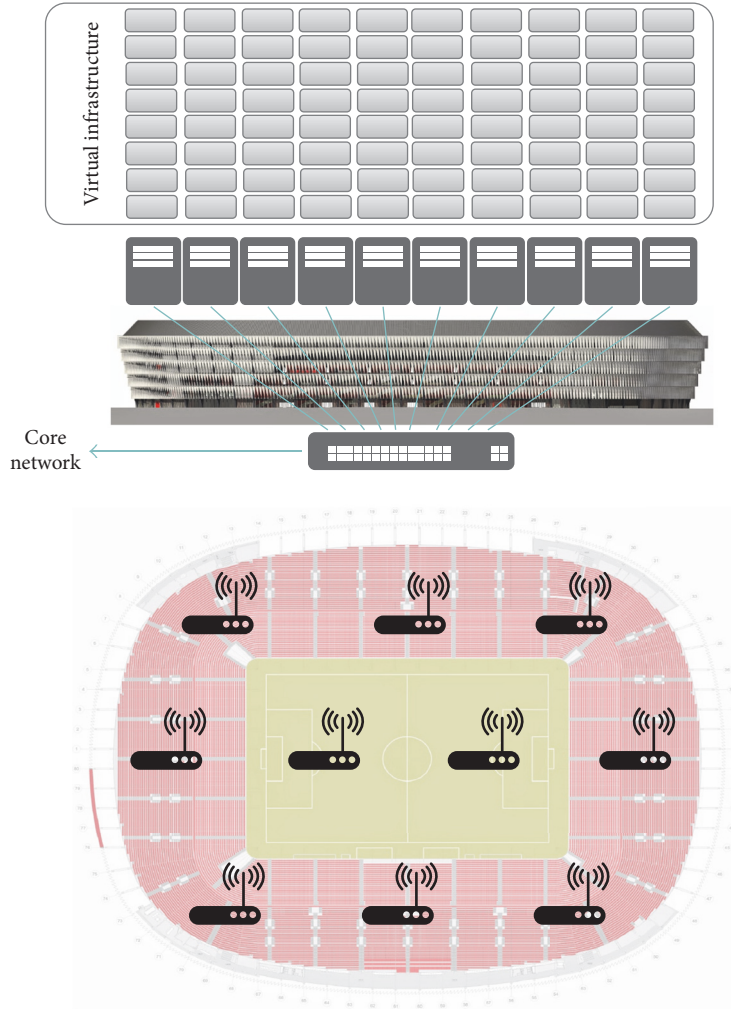
FIGURE 4: Evaluation scenario.

the solving time increases to hours, depending on the specific configuration of the experiment.

Next, we first describe the analysis scenario and the configuration of the performed experiments, to later discuss the results of those experiments. The objective of the experiments is to demonstrate that the placement mechanism is able to allocate multitenant service chains to the available resources considering latency constraints. For this reason, the numerical values employed in the next section are just experimental and do not respond to real world measurements.

5.1. Description of the Analysis Scenario. Our evaluation scenario is depicted in Figure 4. We consider a 10-SC deployment, for example, to cover a football stadium for periodical flash events. The SCs are connected in a star topology through an Ethernet switch. Each SC is composed of 8 cores, 16 GB RAM, 100 GB of storage, and 1 HWA (e.g., GPU).

We also consider 2 NSs that can be offered at 3 service flavours (bronze, silver, and gold) that determine the latency constraint for each NS and the nominal aggregate user bit rate served, as shown in Figure 5.

Table 1 shows the experimental energy model and Table 2 shows the modelling of the VNFs that compose the proposed NSs for the user traffic demand that will be employed in the evaluation experiments. The user traffic values include the nominal bit rates of services NS1 and NS2 (130 Mbps and 90 Mbps, resp.), as well as the values when a 20% protection parameter is applied to the nominal values (156 Mbps and 108 Mbps, resp.) in order to study the effect of robust optimization techniques on the placement decision.

As a final remark, the placement behavior of VNFs can be altered by the use of HWA. The assignment of a HWA to a VNF implies that the service latency of the correspondent instance is reduced 3 times, at a cost of incrementing the power consumption by 10. It is important to note that when the VNF instance needs more than 1 CPU for executing without the assistance of a HWA, then, if the placement algorithm assigns it to a HWA, the VNF will need a single core.

5.2. Discussion of the Placement Results. As stated before, the evaluation scenario includes 3 tenants that hire a combination

| NS1 | VNF1 | VNF2 | VNF3 | VNF5 | VNF6 |
| NS2 | VNF1 | VNF2 | VNF3 | VNF4 | |

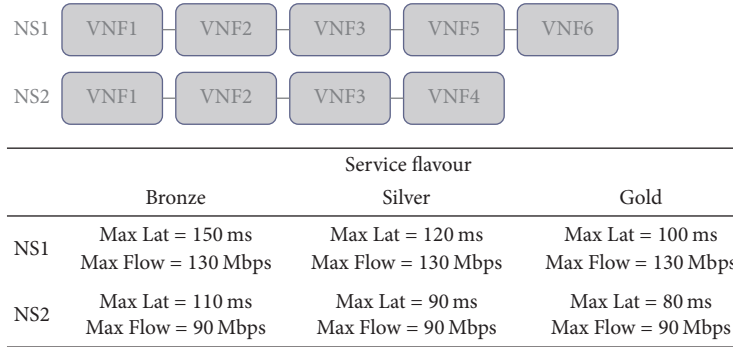| | Service flavour | | |
	Bronze	Silver	Gold
NS1	Max Lat = 150 ms Max Flow = 130 Mbps	Max Lat = 120 ms Max Flow = 130 Mbps	Max Lat = 100 ms Max Flow = 130 Mbps
NS2	Max Lat = 110 ms Max Flow = 90 Mbps	Max Lat = 90 ms Max Flow = 90 Mbps	Max Lat = 80 ms Max Flow = 90 Mbps

FIGURE 5: Description of NSs and service levels.

TABLE 1: Energy models of the cores of the microservers and of the network switch related to the usage percentage.

Usage	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
P_{router}	0	1	2	5	7	9	11	13	14	15	15
P_{cpu}	5	50	70	100	130	150	165	170	180	185	190

TABLE 2: VNF characterization for the service flavours of the evaluation scenario.

	VNF1	VNF2	VNF3	VNF4	VNF5	VNF6
Number of cores per instance						
90 Mbps	1	2	1	1	1	1
108 Mbps	1	2	2	1	1	1
130 Mbps	1	2	2	1	1	1
156 Mbps	1	2	2	1	1	1
CPU usage						
90 Mbps	30%	20%	90%	50%	50%	20%
108 Mbps	40%	30%	60%	60%	60%	30%
130 Mbps	40%	30%	60%	60%	60%	30%
156 Mbps	40%	30%	90%	40%	20%	30%
Service latency						
90 Mbps	10 ms	15 ms	30 ms	35 ms	20 ms	25 ms
108 Mbps	10 ms	15 ms	35 ms	40 ms	25 ms	30 ms
130 Mbps	10 ms	15 ms	35 ms	40 ms	25 ms	30 ms
156 Mbps	12 ms	16 ms	40 ms	50 ms	30 ms	35 ms

of network services/flavours to the SC operator. In this way, in the following, first we are going to evaluate the behavior of the placement algorithm for the three service levels, and next we will show the effect of applying RO techniques to those results.

5.2.1. Bronze Flavour Scenario Results with 0% Robust Protection. Figure 6 shows the placement results of an scenario where all the network services operate at bronze flavour. Different colours represent each tenant in order to differentiate the VNFs that belong to its NS, and different tones distinguish NSs of the same tenant. As a result of the placement algorithm, all the NSs are allocated in the available virtualized resources with energy consumption cost of 3712 W, meeting all the latency constraints. The figure

displays the assignment of VNFs to the SCs, highlighting some placement particularities. As explained in Section 3, the VNF chains of the same tenant share their SCVNF. In our example, Tenant 1 and Tenant 2 both have two NSs that are served with a single SCVNF. Furthermore, each CPU executes a single VM, but some VNF instances need two cores to manage the user demand. In connection with this behavior, the VNF scalability feature is also visible in Figure 6, observing the placement of, for example, VNF3 for the different NSs. NS1, with a higher aggregated user demand needs two CPUs to instantiate VNF3, but since NS2 operates at a lower user demand, the instantiation of VNF3 needs just a single core. Note, as well, that bronze service flavour does not require HW acceleration to meet the latency constraints of the NSs. Finally, it can also be observed that
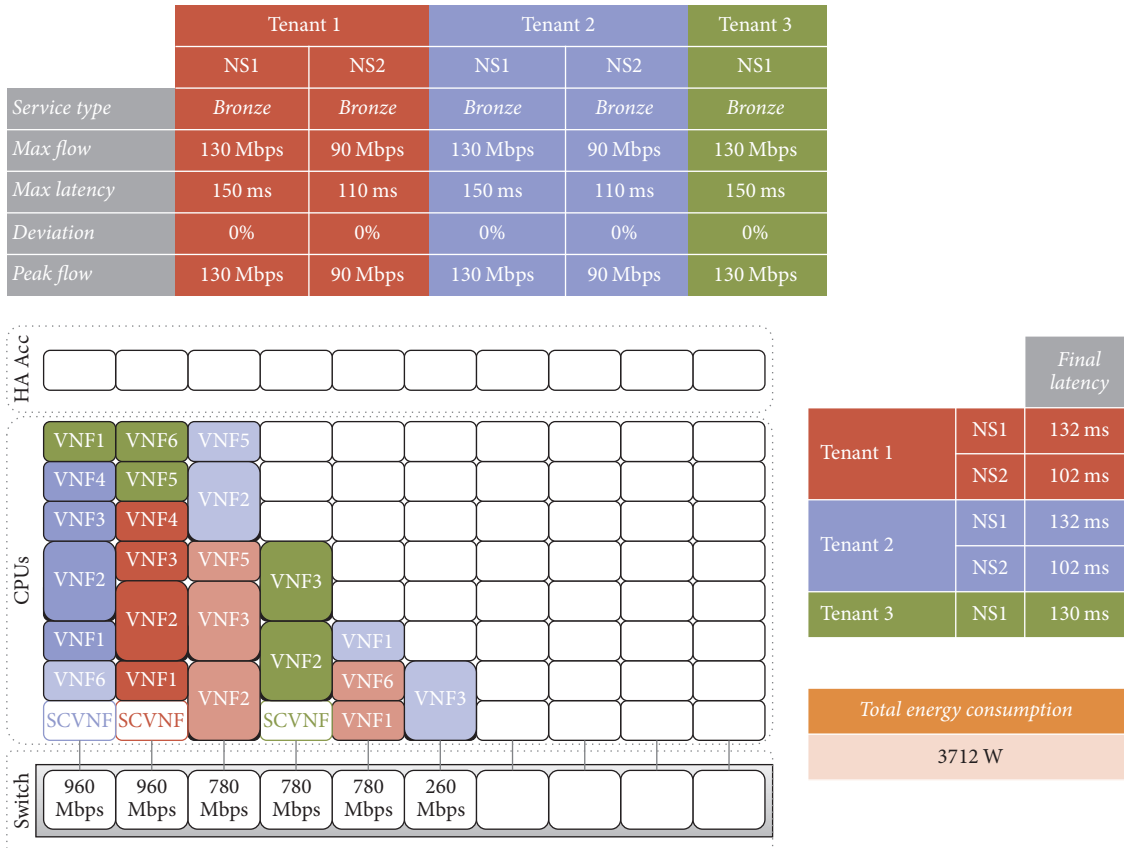
	Tenant 1		Tenant 2		Tenant 3
	NS1	NS2	NS1	NS2	NS1
Service type	*Bronze*	*Bronze*	*Bronze*	*Bronze*	*Bronze*
Max flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps
Max latency	150 ms	110 ms	150 ms	110 ms	150 ms
Deviation	0%	0%	0%	0%	0%
Peak flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps

FIGURE 6: Placement results of bronze flavour setting with 0% robust protection.

		Final latency
Tenant 1	NS1	132 ms
	NS2	102 ms
Tenant 2	NS1	132 ms
	NS2	102 ms
Tenant 3	NS1	130 ms

Total energy consumption
3712 W

the algorithm tries to group as many VNFs as possible in each server, in order to reduce the data traffic across the switch and minimize its load and, therefore, its energy waste.

5.2.2. Silver Flavour Scenario Results with 0% Robust Protection. The evaluation experiment analyzed in this section studies the outcome of the placement algorithm when the latency constraints become more strict with silver service flavour. Due to the further limitation of maximum latency values (20% stricter), the placement mechanism is compelled to use HW acceleration units. Figure 7 shows that, at this service level, all the NSs must use HW acceleration in one of the VNFs composing the service chain. As a consequence, the resulting total latency of the NSs decreases to meet the service level requirements, but, in return, the global power consumption grows by 22% up to 4547 W.

Apart from that, the placement results illustrate the same behavior as the bronze flavoured example of Section 5.2.1 regarding the allocation of VNFs in the available resources, scale-in/out features, and multitenancy.

The evaluation experiments of Sections 5.2.1 and 5.2.2 operate with a nominal aggregated user traffic to perform the optimization decision-making process. However, a perfectly known user traffic demand is seldom available. Usually, there is a certain level of uncertainty in the behavior of the user demand. We include robust optimization techniques to allow the placement algorithm to insert a defined level of uncertainty and protect the system against the undesirable effects of demand peaks. The evaluation examples in Sections 5.2.3 and 5.2.4 include a robust protection parameter and analyze its effect on the global power consumption.

5.2.3. Silver Flavour Scenario Results with 20% Robust Protection. This section studies the placement outcome of the evaluation scenario described in the previous section introducing a 20% robust protection parameter. This modification on the scenario configuration implies an increase of the 20% of the aggregated user traffic as an input of the decision-making process in order to be prepared for potential demand peaks during the service operation.

Figure 8 exhibits that VNF3 of NS2 must scale as a consequence of the traffic growth and now needs two CPUs for each instance. The traffic increase also has an impact on the processing latency of the VNFs causing an adjustment in the use of HW accelerators. As a consequence, some VNFs that required 2 CPUs for their operation in the previous evaluation scenario now only need a single core in combination with the HWA. Therefore, in some cases, the higher power consumption that is implicit to the use of HWAs may be balanced with a decrease on the number of CPUs required for the VNF. Besides, the higher aggregated user traffic involves a more intense use of the network switch. These factors have

	Tenant 1		Tenant 2		Tenant 3
	NS1	NS2	NS1	NS2	NS1
Service type	*Silver*	*Silver*	*Silver*	*Silver*	*Silver*
Max flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps
Max latency	120 ms	90 ms	120 ms	90 ms	120 ms
Deviation	0%	0%	0%	0%	0%
Peak flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps

		Final latency
Tenant 1	NS1	100 ms
	NS2	88 ms
Tenant 2	NS1	118 ms
	NS2	88 ms
Tenant 3	NS1	116 ms

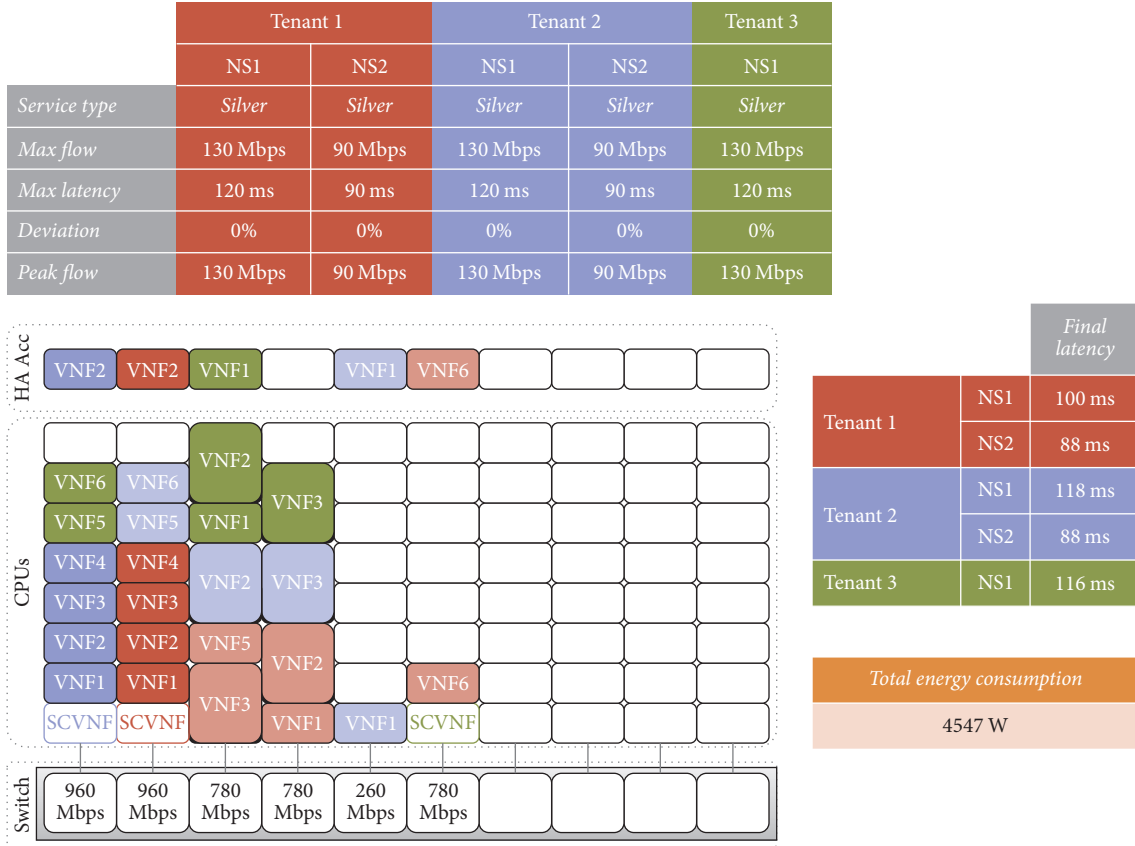| *Total energy consumption* |
|---|
| 4547 W |

Figure 7: Placement results of silver flavour setting with 0% robust protection.

an impact on the power consumption of the components of the network, resulting in a global energy consumption of 5047 W, which implies an increment of 11% compared to the unprotected placement.

5.2.4. Gold Flavour Scenario Results with 20% Robust Protection. Finally, the last evaluation example sets even stricter latency constraints to the NSs to be placed and includes a 20% robust protection parameter. Figure 9 depicts the results of the placement algorithm with this configuration. Again, the reduction of latency values forces the increment of the use of HW acceleration. In this case, the main consequence is the spreading of the VNFs across the SCs to use the single HWA available in each cell. The use of HWAs, the increase of power consumption due to higher user aggregated traffic, and the higher network traffic traversing the switch cause the global energy consumption to grow up to 5569 W.

As a final remark, Figure 10 shows the comparison of the energy consumption and use of resources (cores, active CESCs, and HWA) related to the robust protection level in a wider collection of scenarios. In particular, the charts of Figure 10 compare the results of the placement algorithm for the three service levels described in the previous sections (bronze, silver, and gold) combined with three robust protection levels (0% or no protection, 10%, and 20%). Obviously, the main trend is that both power and resource consumption increase with higher values of protection, but the associated cost can be assumed in favor of reducing the impact of

unexpected user demand peaks. However, there are some cases in which the protection parameter does not have so adverse implications. For example, it can be observed how the number of used cores in the silver service level with a 10% protection level is lower than in the case of no robust protection. The reason of this behavior is how the algorithm uses HWA appliances in this case. The increase of the user traffic demand in the protected services implies the scaling up of some VNFs that now must use more cores for the operation of the NSs. Then, the placement algorithm decides to use HW accelerators for those specific VNFs, compensating this way the higher energy cost of HWAs with a lower use of CPU cores. With all these considerations that can be extracted from a deeper analysis of robust placement results robust optimization can be applied to decision-making problems, not ending just with the optimization problem.

The proposed steps in this paper can be used in any placement decision-making context and its goal is to provide a tool that helps system administrator to evaluate the possible impact of any decision on the performance of the deployment underneath.

6. Conclusions

This paper introduces a constraint-based expert system to support the decision-making process of VNF chain placement in distributed edge cloud networks. The placement

	Tenant 1		Tenant 2		Tenant 3
	NS1	NS2	NS1	NS2	NS1
Service type	*Silver*	*Silver*	*Silver*	*Silver*	*Silver*
Max flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps
Max latency	150 ms	110 ms	150 ms	110 ms	150 ms
Deviation	20%	20%	20%	20%	20%
Peak flow	156 Mbps	108 Mbps	156 Mbps	108 Mbps	156 Mbps

		Final latency
Tenant 1	NS1	116 ms
	NS2	79 ms
Tenant 2	NS1	110 ms
	NS2	79 ms
Tenant 3	NS1	118 ms

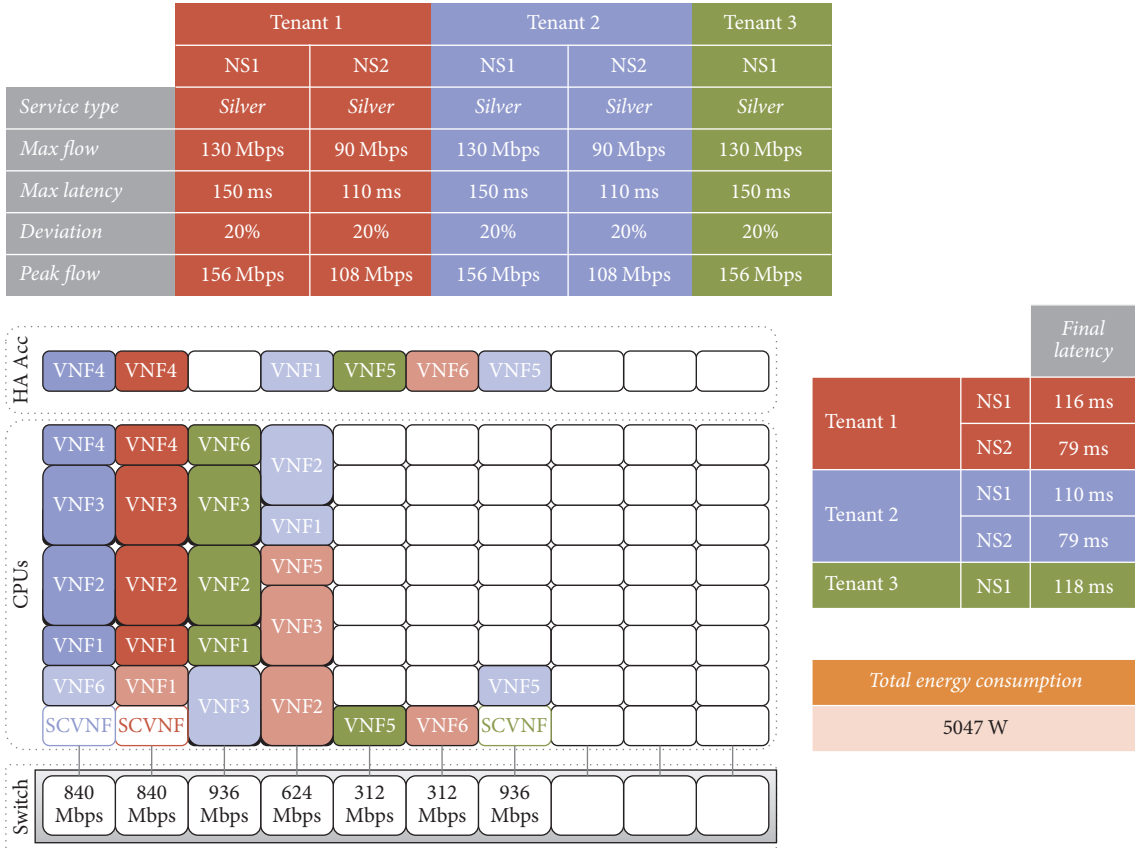| *Total energy consumption* |
|---|
| 5047 W |

FIGURE 8: Placement results of silver flavour setting with 20% robust protection.

algorithm applies robust optimization techniques to the minimization of the global power consumption of the system, subject to the bit rate and latency constraints of the network services. The power consumption of the components of the system depends on the usage percentage, which is a consequence of the aggregated user traffic demand. The model assumes that the relationships between the user traffic demand, the power consumption, and the use of the resources may be nonlinear. Besides, the placement algorithm considers the 5G communication particularities of both control and data plane and allows multitenancy. Finally, the algorithm also includes VNF scaling features and the use of hardware acceleration.

The results show that the placement algorithm succeeds in allocating the VNFs that compose the NSs of different tenants in the available resources, meeting the latency constraints and minimizing the energy consumption of the whole system. The presented evaluation examples demonstrate the features of the constraint-based expert system and illustrate the impact of the selected service flavour and robust protection on the global power consumption and, therefore, the exploitation cost of the 5G distributed edge cloud.

Further research will extend the characterization of VNFs to include variable output bit rate. This feature will model the behavior of, for example, transcoding functions. A future version of the algorithm will also incorporate the existence of a virtual switch to connect the VMs running in the same microserver. Another enhancement of the model will be the combination of heterogeneous SCs with different configurations and resource consumption models. Finally, the obtained placement results will be used for a technoeconomic study that will analyze the relationship between the server aggregated user traffic cost of the required infrastructure and the pricing of the services. This study will also consider the integration of the placement decision problem with the radio side.

Summary of Model Indexes, Sets, Parameters, and Decision Variables

Indexes

π: Index for VNF placement policy

x: Index for SC microserver

y: Index for core number in each microserver

i: Index for NS

j: Index for VNF in a network service chain
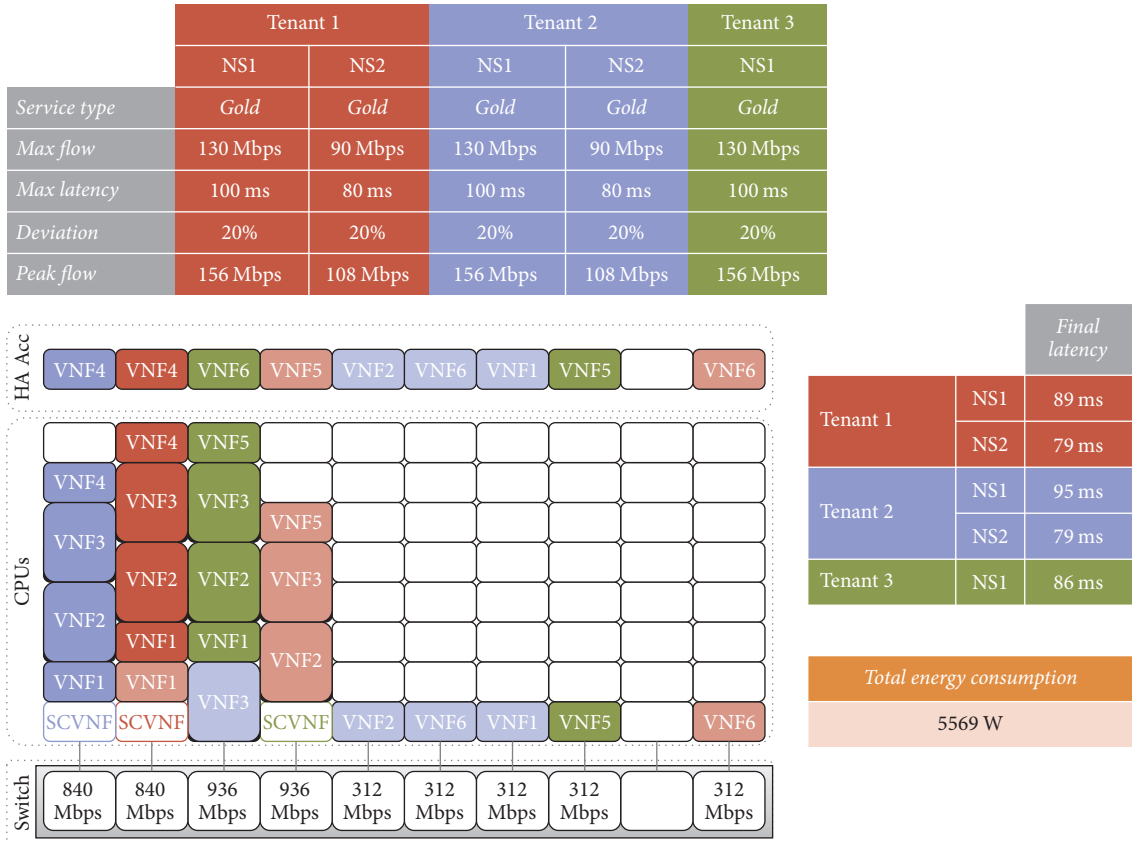
r: Index for physical resources in SC microserver.

	Tenant 1		Tenant 2		Tenant 3
	NS1	NS2	NS1	NS2	NS1
Service type	*Gold*	*Gold*	*Gold*	*Gold*	*Gold*
Max flow	130 Mbps	90 Mbps	130 Mbps	90 Mbps	130 Mbps
Max latency	100 ms	80 ms	100 ms	80 ms	100 ms
Deviation	20%	20%	20%	20%	20%
Peak flow	156 Mbps	108 Mbps	156 Mbps	108 Mbps	156 Mbps

		Final latency
Tenant 1	NS1	89 ms
	NS2	79 ms
Tenant 2	NS1	95 ms
	NS2	79 ms
Tenant 3	NS1	86 ms

Total energy consumption
5569 W

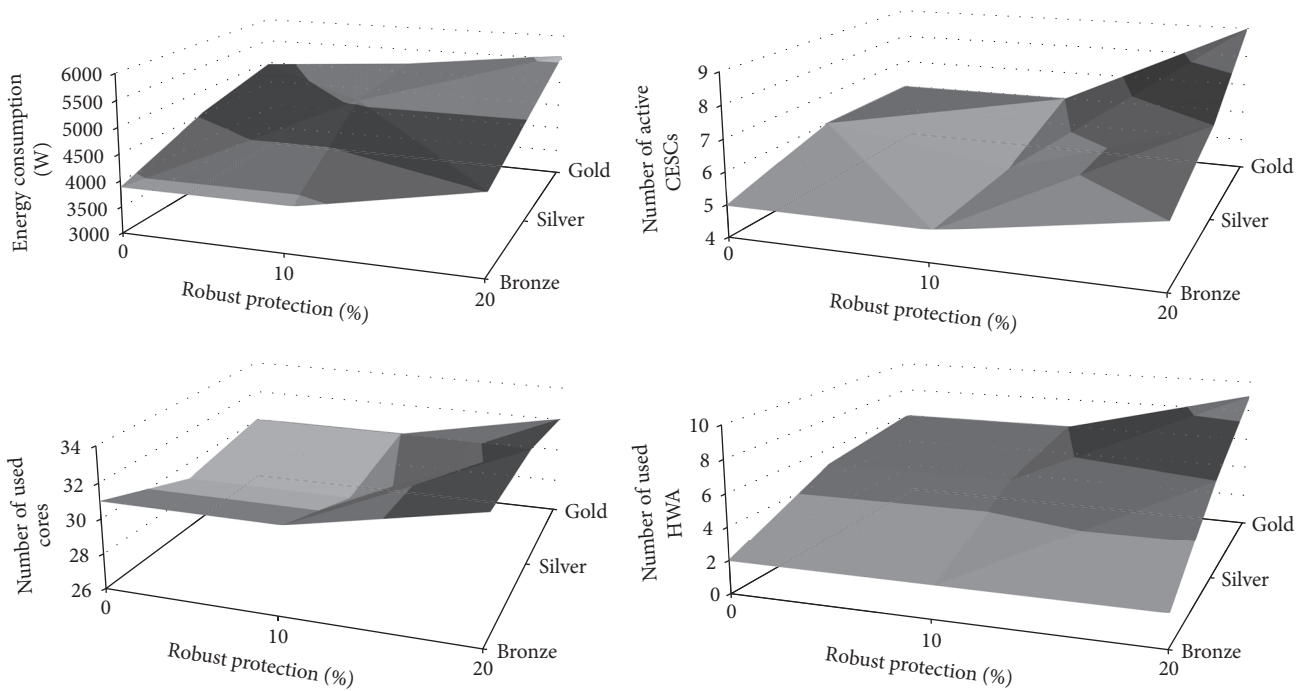FIGURE 9: Placement results of gold flavour setting with 20% robust protection.

FIGURE 10: Comparison of energy and resource consumption for different levels of robust protection.

Sets

Π: Set of possible VNF placement policies
X: Set of SC microservers
Y: Set of cores composing a SC microserver
I: Set of offered NSs
J_i: Set of VNFs in NS i
R: Set of resources available in
each SC microserver.

Input Parameters

T_i: Contracted maximum aggregated user traffic for network service i
$D_{\max_i}$: Maximum latency allowed for NS i
M_{rx}: Amount of resource r available in SC microserver x
Psc_x: Power consumption of SC x because of being switched on
$Pcpu_{ij}^{xy}$: Power consumption of VM y of SC x due to CPU use of VNF j of NS i
$Pswitch_{ij}$: Power consumption in the physical switch due to CPU use of VNF j of NS i.

Decision Variables

P^π: Total power consumption of placement policy π
P_{ij}^π: Power consumption of VNF j of NS i in placement policy π
m_{rx}^π: Consumption of resource r in SC microserver x in placement policy π
tl_x^π: Traffic of the link that connects SC microserver x with the switch inplacement policy π
A_{ij}^{xy}: Binary variable that expresses the assignment of VNF j of NS i to VM y in SC x
L_{ij}: Binary variable to express the use of the network switch to forward the flow of VNF j of NS i to the next VNF
d_i: Delay of NS i
d_{ij}: Delay of VNF j of NS i
dp_{ij}^{xy}: Processing delay in the CPU of VNF j of NS i in VM y of SC x
$dl_{ij}^{x_1 y_1, x_2 y_2}$: Path delay for NS i between VNF j (located in VM y_1 of SC x_1) and VNF $j + 1$ (located in VM y_2 of SC x_2)
$d_{vs,xy}$: Link delay between the virtual switch in SC x and the VM y in SC x
$d_{vs_x,s}$: Link delay between the virtual switch in SC x and the physical switch
d_x^{vs}: Delay in the virtual switch of SC x
d_s: Delay in the physical switch.

Conflicts of Interest

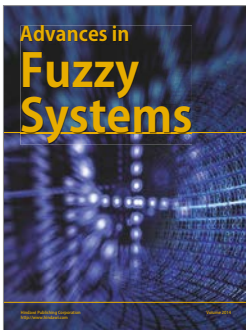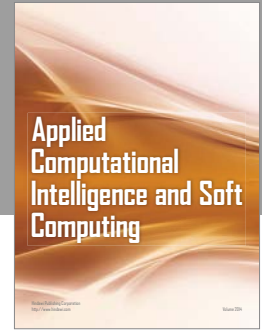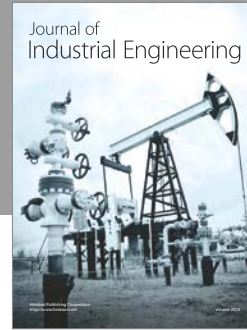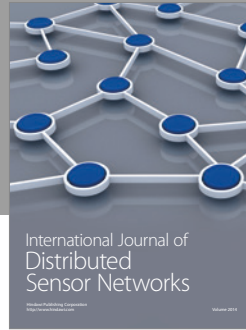The authors declare that they have no conflicts of interest.

References

[1] ITU-R, M.2083: IMT Vision-Framework and overall objectives of the future development of IMT for 2020 and beyond, 2015.

[2] P. Demestichas, A. Georgakopoulos, D. Karvounas et al., "5G on the Horizon: key challenges for the radio-access network," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 47–53, 2013.

[3] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17–48, 2016.

[4] ETSI, Mobile-edge computing: Introductory technical white paper, 2014.

[5] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[6] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: state-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[7] Ericsson White Paper, Cloud-RAN– the benefits of virtualization, centralization and coordination, 2015.

[8] B. Blanco, J. O. Fajardo, I. Giannoulakis et al., "Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN," *Computer Standards and Interfaces*, vol. 54, pp. 216–228, 2017.

[9] ETSI, Network Functions Virtualization (NFV), Management and Orchestration, 2014.

[10] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*, Aharon Ben-Tal, Laurent El Ghaoui, Arkadi Nemirovski-Google Books, Princeton University Press, Princeton, New Jersey, NY, USA, 2009.

[11] J. Gil Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[12] B. Addis, D. Belabed, M. Bouet, and S. Secci, "Virtual network functions placement and routing optimization," in *Proceedings of the 4th IEEE International Conference on Cloud Networking, CloudNet '15*, pp. 171–177, October 2015.

[13] H. Moens and F. De Turck, "VNF-P: a model for efficient placement of virtualized network functions," in *Proceedings of the 10th International Conference on Network and Service Management (CNSM '14)*, pp. 418–423, Rio de Janeiro, Brazil, November 2014.

[14] A. Gupta, M. F. Habib, P. Chowdhury, M. Tornatore, and B. Mukherjee, "On service chaining using Virtual Network Functions in Network-enabled Cloud systems," in *Proceedings of the 9th IEEE International Conference on Advanced Networks and Telecommunications Systems, ANTS '15*, pp. 1–3, December 2015.

[15] M. Bouet, J. Leguay, T. Combe, and V. Conan, "Cost-based placement of vDPI functions in NFV infrastructures," *International Journal of Network Management*, vol. 25, no. 6, pp. 490–506, 2015.

[16] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle:

efficient placement and chaining of virtual network functions," in *Proceedings of the 14th International Symposium on Integrated Network Management, IM '15*, pp. 98–106, IEEE, Toronto, Canada, May 2015.

[17] S. Kim, Y. Han, and S. Park, "An energy-Aware service function chaining and reconfiguration algorithm in NFV," in *Proceedings of the 1st International Workshops on Foundations and Applications of Self-Systems, FAS-W '16*, pp. 54–59, September 2016.

[18] V. Eramo, A. Tosti, and E. Miucci, "Server resource dimensioning and routing of service function chain in NFV network architectures," *Journal of Electrical and Computer Engineering*, vol. 2016, Article ID 7139852, 12 pages, 2016.

[19] K. Hida and S.-I. Kuribayashi, "Virtual routing function allocation method for minimizing total network power consumption," *World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, vol. 10, pp. 997–1002, 2016.

[20] C. Pham, N. H. Tran, S. Ren, W. Saad, and C. S. Hong, "Traffic-aware and energy-efficient vnf placement for service chaining: joint sampling and matching approach," *IEEE Transactions on Services Computing*, 2017.

[21] A. Marotta and A. Kassler, "A power efficient and robust virtual network functions placement problem," in *Proceedings of the 28th International Teletraffic Congress, ITC '16*, pp. 331–339, September 2016.

[22] C. Pham, H. D. Tran, S. I. Moon, K. Thar, and C. S. Hong, "A general and practical consolidation framework in CloudNFV," in *Proceedings of the 2015 International Conference on Information Networking, ICOIN '15*, pp. 295–300, IEEE, Cambodia, Cambodia, January 2015.

[23] V. Eramo, M. Ammar, and F. G. Lavacca, "Migration energy aware reconfigurations of virtual network function instances in NFV architectures," *IEEE Access*, vol. 5, pp. 4927–4938, 2017.

[24] N. E. Khoury, S. Ayoubi, and C. Assi, "Energy-aware placement and scheduling of network traffic flows with deadlines on virtual network functions," in *Proceedings of the 5th IEEE International Conference on Cloud Networking, CloudNet '16*, pp. 89–94, IEEE, Pisa, Italy, October 2016.

[25] V. Eramo, E. Miucci, M. Ammar, and F. G. Lavacca, "An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures," *IEEE/ACM Transactions on Networking*, 2017.

[26] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Energy-efficient robust resource provisioning in virtualized wireless networks," in *Proceedings of the IEEE International Conference on Ubiquitous Wireless Broadband, ICUWB '15*, pp. 1–5, IEEE, Montreal, Canada, October 2015.

[27] S. Coniglio, A. Koster, and M. Tieves, "Data uncertainty in virtual network embedding: robust optimization and protection levels," *Journal of Network and Systems Management*, vol. 24, no. 3, pp. 681–710, 2016.

[28] I. Takouna, K. Sachs, and C. Meinel, "Multiperiod robust optimization for proactive resource provisioning in virtualized data centers," *Journal of Supercomputing*, vol. 70, no. 3, pp. 1514–1536, 2014.

[29] G. Chochlidakis and V. Friderikos, "Robust virtual network embedding for mobile networks," in *Proceedings of the 26th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC '15*, pp. 1867–1871, IEEE, Kowloon, Hong Kong, September 2015.

[30] S. Coniglio, A. M. Koster, and M. Tieves, "Virtual network embedding under uncertainty: exact and heuristic approaches," in *Proceedings of the 11th International Conference on the Design of Reliable Communication Networks, DRCN '15*, pp. 1–8, IEEE, Kansas, Kan, USA, March 2015.

[31] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, vol. 52, no. 1, pp. 35–53, 2004.

[32] F. Rossi, P. Van Beek, and T. Walsh, *Handbook of Constraint Programming*, Elsevier, Amsterdam, Netherland, 2006.

[33] http://www.minizinc.org/.

[34] C. Schulte, G. Tack, and M. Z. Lagerkvist, *Modeling and programming with gecode*, 2010, Schulte, Christian and Tack, Guido and Lagerkvist, Mikael.

Advances in
Multimedia

The Scientific
World Journal

International Journal of
Distributed
Sensor Networks

Journal of
Industrial Engineering

Applied
Computational
Intelligence and Soft
Computing

Advances in
Fuzzy
Systems

Modelling &
Simulation
in Engineering

Journal of
Computer Networks
and Communications
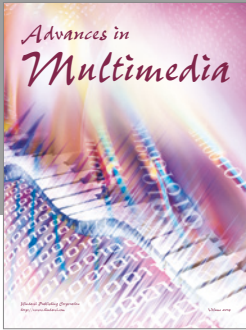
Hindawi

Submit your manuscripts at
https://www.hindawi.com

Advances in
Artificial
Intelligence

Advances in
Computer Engineering

International Journal of
Computer Games
Technology

International Journal of
Biomedical Imaging

Advances in
Artificial
Neural Systems

Advances in
Software Engineering

Journal of
Robotics

Advances in
Human-Computer
Interaction

Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering