Journal of Network and Computer Applications 114 (2018) 29–37

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Latency-aware cost optimization of the service infrastructure placement in 5G networks

Alejandro Santoyo-González *, Cristina Cervelló-Pastor

Department of Network Engineering, Universitat Politècnica de Catalunya (UPC), Esteve Terradas, 7, 08860 Castelldefels, Spain

ARTICLE INFO

Keywords:
5G
NFV
Fog Computing
Optimization
MILP

ABSTRACT

Under 5G use case scenarios latency is a main challenge that must be addressed, since mission critical environments are mostly delay sensitive. To achieve this goal, the service infrastructure placement optimization is needed in the interest of minimizing the delays in the service access layer. To solve this problem, this paper mathematically models the placement problem in a Fog Computing/NFV environment as a Mixed-Integer Linear Programming problem and proposes a heuristic-based solution considering 5G mobile network requirements. As a practical result, an application was developed to achieve usability and flexibility while ensuring operational applicability of the proposed methods.

1. Introduction

Albeit some remaining skepticism related to whether identified 5G requirements could be satisfied for all use case scenarios, an upcoming communication revolution based on 5G is a foreseeable future. In fact, the 5G roadmap envisions the first operational networks by 2021 (Rodriguez, 2015). Under these circumstances, technologies such as Network Function Virtualization (NFV), Fog Computing (FC) and Software Defined Networking (SDN) will most likely converge as pillars to answer underlying technological challenges.

In 5G usage scenarios, latency is certainly one of the main difficulties to overcome. The scenario classification devised by the International Telecommunications Union-Radiocommunication Sector (ITU-R), shows mission-critical services depending on strong latency constraints (Xiang et al., 2017). Specifically, these latency constraints are expected to reach less than 1 ms in extremely demanding use cases such as autonomous driving (Xiang et al., 2017). Given this situation, *latency control* and *latency-aware planning* become mandatory.

A first solution approach to this problem, is to minimize the delay in the service location-end user location channel by placing the service infrastructure at the network edge and optimizing the placement strategies. This target could be achieved through NFV and FC convergence supported by a high performance and carefully designed network (presumably an SDN solution). Therefore, deploying small-sized infrastructure *containers* (fog nodes in this context) over the service area would lead to a significant reduction in service access delay values.

Despite the existence of a pointer towards the right direction, capital and operational expenditures (CAPEX and OPEX) remain significant challenges operators have to face in 5G networking. Infrastructure capacities, for instance, will certainly become a trade-off to rethink, since 5G requirements will pose a complex challenge to cost-efficiency and the avoidance of under-utilization will remain a critical issue. In addition, as there are few studies and none working knowledge in the deployment and operation of infrastructure/services over 5G ecosystems, there is no cost-effective general method to translate use case demands into infrastructure capacities. Thus, reducing CAPEX and OPEX in the context of this paper, is targeted through the minimization of the fog nodes number and their capacities.

Mobile network planning studies and facility location research have exhaustively addressed similar problems (Wang et al., 2015; Barbati, 2013; Zhou et al., 2015; Carlsson et al., 2016; Wang and Ran, 2016). Nevertheless, the present research is an ongoing work targeting the particular ecosystem merging 5G, NFV, FC and, in the future, SDN. Furthermore, it aims to reduce deployment costs by minimizing both the number of facilities deployed, and the location-dependent expenses such as building costs and operating costs.

For these reasons, the following considerations were made:

- Initial facility coordinates are co-locations of identified *traffic generators* (TG).
- The fog node (FN) *coverage area* is determined as a function of an assumed latency value according to 5G requirements.

* Corresponding author.
E-mail addresses: alejandro.santoyo@entel.upc.edu (A. Santoyo-González), cristina@entel.upc.edu (C. Cervelló-Pastor).
https://doi.org/10.1016/j.jnca.2018.04.007
Received 15 October 2017; Received in revised form 16 February 2018; Accepted 15 April 2018
Available online 17 April 2018

Although FC has been thoroughly studied and remains under active research, to the best of our knowledge there is very little or no available literature targeting the FN placement problem (FNPP). Authors in (Intharawijitr et al., 2016) proposed a mathematical formulation analyzing FC communication concerns regarding latency and workload intensity. The proposal covers a scenario where both communication delay and processing latency are considered, while the main goal is to find a suitable FN set complying with the specified restrictions. At this point, a predefined policy set allows the system to select an appropriate FN to process a given workload. In general, this article does not tackle the FN site selection problem, where a main issue is the communication delay between the data sources and the FNs. On the other hand, the bipartite graph representation proposed does not accurately match future FC scenarios, as in such environment, service tasks are to be shared and cooperatively executed among both the FN set and the end-user smart devices. In (Balevi and Gitlin, 1801) the optimum number of FNs is using stochastic geometry analysis by representing user demands through Binomial Point Process. The idea is to select, from a deployed network of nodes, which of them are going to be upgraded to FN based on a certain probability. The overall aim is to reduce the total number of upgraded nodes since this yields better signal-to-interference-plus-noise ratio (SINR), average data rate and transmission delay.

In (Yannuzziet al, 2017) the vendor-specific infrastructure deployment at the edge of the network is studied and the lack of flexibility of such scenario (for next generation networks) is pointed out. FC is then considered as a solution to the defined problem and street cabinets are proposed as suitable locations to deploy the FNs for a particular case (the city of Barcelona). This site selection is extended to the notion of points-of-presence in more general deployments. From such proposal, a first approach to the FNPP could be extrapolated, as the intention is to find suitable locations to place the service infrastructure close enough to the data sources.

As of the crucial relevance of latency in fog scenarios, from (Byers, 2017), 7 of the 12 presented use cases (only considering IoT-related vertical sectors), are said to be latency-constrained. Similarly, in a broader scope analysis, thus including 5G scenarios (Blancoet al, 2017), and (Elayoubiet al, 2016) offer clear evidence of the critical role played by latency control and optimization.

There are other relevant studies targeting FC-based configurations where latency is a main concern. Regarding resource allocation and capacity sizing in the context under analysis (Basta et al., 2017), presents a cost optimization model where the network load is proposed as optimization target and defined as the bandwidth-latency product. Likewise, in (Brogi and Forti, 2017), the resource allocation problem in FC is solved considering bandwidth and latency as core parameters. In this article, although an operational and thus already deployed FC infrastructure is considered known data (implicitly the FNPP is previously solved and its solution assumed as input), a novel approach presented is to model the allocation problem considering the possible interconnecting and interaction methods in a fairly realistic architecture. Gravalos et al. (2016) formulate the problem of deploying IoT gateways as an Integer Linear Programming (ILP). They aim to reduce the costs of the IoT network under a strict latency constraint. Their formulation achieves interesting cost reduction results, although the number of nodes used in the simulations is to low to represent a real-life scenario. What is more, such formulation is limited by the exponential complexity derived from scaling the number of devices and gateways required.

From (Zhang et al., 2017) a first solution to the node site selection problem under FC is solved. The problem presented is the location selection to place micro datacenters and diverse components of a long-reach passive optical network such that the total cost of deployment is reduced. The problem is formulated as an ILP and solved using an in-house developed heuristic. A limitation of this study is the isolated nature of the deployed micro datacenters. The proposed model prohibits micro datacenters collaboration for service execution. Consequently, the solutions found could be under sized, as in a shared and cooperative environment more datacenters should be deployed. In addition, future FN or even (broadening the scope) edge nodes (when considering other technologies such as Mobile Edge Computing and Cloudlets (Dolui and Datta, 2017)) are most likely to converge a wide range of interconnecting technologies merging highly dense service areas (from pico to macro-cells).

The so-called location selection problem, has also been extensively researched and is commonly denoted Facility Location Problem (FLP). Usually, FLPs deal with selecting the placement of a facility or a set of facilities (often from a list of feasible locations) to best meet the demanded constraints and user requirements. Their solution is quite useful when planning the placement of public service facilities such as hospitals, fire fighter stations or commercial facilities such as warehouses. In (Arabani and Farahani, 2012) an extensive survey about this topic can be found.

Traditional FLP formulations cannot be directly applied to the FNPP mainly because the FNPP treated here, do not converge into a particular problem type. As seen in (Farahani et al., 2010), FLPs are mostly formulated following the guidelines of an operational research family: Weber, median, covering, constrained, uncapacitated, location–allocation, location-routing, dynamic, competitive, network and undesirable location problems. Therefore, they all converge into a particular problem type such as coverage or Weber, while the FNPP mixes more than one problem family. Moreover, another difference with the FNPP is that FLPs do not consider non-technical restrictions in the location selection process. Our proposal is to consider the FNPP as a capacity and latency constrained location selection problem.

This formulation converges to a covering problem where the added capacitated approach introduces additional complexity and prohibits the use of the models and solution methods found in the reviewed literature. In particular, the limited capacity and latency restrictions prevented us from using traditional local search techniques (Krivitski et al., 2005). This is firstly because any “move” made to generate a new feasible solution could lead to a worse overall cost, thus forcing us to escape local optima. What is more, deselecting a location usually forces an adjustment procedure in the whole deployment to ensure that all TGs are covered.

In the context mentioned before, latency-aware planning becomes vital for the FN efficient deployment and user demand satisfaction in current and 5G envisioned scenarios. Therefore, the **objective** of this paper is to propose a cost-effective latency-aware strategy to the FN placement in 5G environments. To this aim, in this document we present the mathematical modeling of the problem, merging both a capacitated and a coverage facility location approach (Arabani and Farahani, 2012; Ulukan and Demircioglu, 2015), along with a solution strategy, where the main contributions are:

- A mathematical formulation (as a linear optimization problem) of the service infrastructure placement problem at the network edge (fog nodes).
- A heuristic approach called Hybrid Simulated Annealing to solve optimization problem, based on the location of the *traffic generators*.
- A desktop application including both solution approaches, in which the following data can be obtained: the FN locations and capacities, the TG covered by each FN, and the number of deployed FNs.

2. Formulation as a MILP problem

In order to model the above mentioned problem, we first assume that the service users distributed over a given area could be modeled as *traffic generators* (TG) (Wang et al., 2015; Wang and Ran, 2016). Such simplification is made considering that the last-mile access infrastructure is envisioned to be wireless for most 5G usage scenarios. Thus, the aggregated cell structure composed by mobile base stations, wireless access points, etc. (macro cells, micro cells, femto cells), is used as base

entry data and these aggregation points are defined as *traffic generators*. As the major objective is to reduce costs by minimizing the number of FNs while considering a limited capacity for each FN, the optimization problem is then formulated as follows:

$$\text{Minimize:} \quad \sum_{\forall f \in FN} v_f \tag{1}$$

$$\text{s. t.:} \sum_{\forall f \in FN} u_{tf} \geq 1 \qquad \forall t \in TG \tag{2}$$

$$u_{tf} = v_f \quad \text{if loc(t) = loc(f)} \qquad \forall t \in TG,\ \forall f \in FN \tag{3}$$

$$u_{tf} \leq v_f \quad \text{if loc(t)} \neq \text{loc(f)} \qquad \forall t \in TG,\ \forall f \in FN \tag{4}$$

$$\sum_{\forall f \in FN} d_{tf} = td_t \qquad \forall t \in TG \tag{5}$$

$$d_{tf} \leq td_t \cdot u_{tf} \qquad \forall t \in TG, \forall f \in FN \tag{6}$$

$$\sum_{\forall t \in TG} d_{tf} \leq c(f) \qquad \forall f \in FN \tag{7}$$

$$c(f) = \begin{cases} 0 & \text{if} \quad \sum_{\forall t \in TG} d_{tf} = 0 \\ A & \text{if} \quad 0 < \sum_{\forall t \in TG} d_{tf} < A \\ B & \text{if} \quad A \leq \sum_{\forall t \in TG} d_{tf} < B \\ C & \text{if} \quad B \leq \sum_{\forall t \in TG} d_{tf} \leq C \end{cases} \qquad \forall f \in FN \tag{8}$$

$$\text{if } u_{tf} = 1 \Rightarrow distance(t,f) \leq D_{\max} \qquad \forall f \in FN, \forall f \in FN \tag{9}$$

$$\text{if } u_{tf} = 0 \Leftrightarrow d_{tf} = 0 \qquad \forall t \in TG, \forall f \in FN \tag{10}$$

$$\text{if } v_f = 0 \Leftrightarrow \sum_{\forall t \in TG} d_{tf} = 0 \qquad \forall f \in FN \tag{11}$$

$$u_{tf}, v_f \text{ binary} \qquad \forall t \in TG, \forall f \in FN \tag{12}$$

$$d_{tf} \geq 0 \in \mathbb{R} \qquad \forall t \in TG, \forall f \in FN. \tag{13}$$

Having that:

u_{tf}: 1 if TG t is served by FN f, 0 otherwise
v_f: 1 if FN f is deployed, 0 otherwise
$c(f)$: FN f capacity
d_{tf}: part of TG t demand served by FN f
td_t: total demand of TG t
$D_{\max}$: maximum allowed distance between a TG and its serving FN
$loc(t)$ or $loc(f)$: location of TG t or FN f

The objective function in Eq. (1), seeks to minimize the number of FNs (v_f). The global aim is to select "good" FN locations in terms of delay, capacity and service load. Furthermore, by adjusting the FN capacity to the covered area demands, we also pursue a low-cost solution.

The first set of restrictions (2) specifies that any given TG t should be covered by one or more FNs. The constraint set (3), refer to the case of FN f co-located at TG t position, while (4) ensures that no FN is placed unless there is a TG to *cover*. In (5), TG t demand should be entirely covered by its serving FN f. On the other hand, (6) defines the part of TG t demand served by FN f, in case the association between t and f exists.

From (7), the summation of the covered TG demands under an FN, should not exceed the FN capacity, which is defined in (8). The linearization of the FN capacities as a piecewise constant function is shown in Subsection 2.1.

To fulfill latency-awareness, the parameter $D_{\max}$ is introduced in (9). This parameter is set as the maximum distance allowed between a TG t and its serving FN, such that a given latency value is not exceeded by the placement strategy. As a consequence, any FN location complies with the particular latency requirements imposed to the planning algorithm. The distance between any pair TG-FN was assumed to be the Euclidean distance, therefore, these implication have been linearized as shown in Subsection 2.2.

The set of restrictions (10) relates the part of the TG t demand served by FN f to the binary variable u_{tf}, which determines if this relationship exists indeed. The same idea is applied on (11), guaranteeing that only deployed FNs cover the corresponding part of TGs demand that are associated to them. Both set of constraints are linearized in Subsection 2.3.

Finally, (12) and (13) are variable-type or domain constraints that specify the type of values the decision variables can take.

2.1. Modeling FN capacities

The capacity of each FN is modeled as a piecewise-constant function of P pieces or sections (with $P = 4$), as shown in (8) (see Fig. 1).

In order to linearize such function, the binary variable $\delta_{if} \forall i \in P, f \in FN$ (a δ value per function section), is introduced to determine which capacity should be selected depending on the sum of the demands covered by FN f. The value of δ_{if} is 1 at the ith section and 0 otherwise. As result, the constraints (14)–(19) are added to the model, where A, B and C are the available FN capacities, being C the maximum and A the minimum value. To obtain the inequalities in (15) and (16), as required in the linearization procedures, the value ϵ is defined as an arbitrary small value.

$$\sum_{\forall t \in TG} d_{tf} \leq C \cdot (1 - \delta_{1f}) \qquad \forall f \in FN \tag{14}$$

$$\sum_{\forall t \in TG} d_{tf} \leq C \cdot (1 - \delta_{2f}) + \delta_{2f} \cdot (A - \epsilon) \qquad \forall f \in FN \tag{15}$$

$$\sum_{\forall t \in TG} d_{tf} \leq C \cdot (1 - \delta_{3f}) + \delta_{3f} \cdot (B - \epsilon) \qquad \forall f \in FN \tag{16}$$

$$\sum_{\forall t \in TG} d_{tf} \geq A \cdot \delta_{3f} \qquad \forall f \in FN \tag{17}$$

$$\sum_{\forall t \in TG} d_{tf} \geq B \cdot \delta_{4f} \qquad \forall f \in FN \tag{18}$$

$$\sum_{\forall t \in TG} d_{tf} \leq C \qquad \forall f \in FN \tag{19}$$

Moreover, variable δ_{if}, $\forall i \in \{1, \ldots, P\}$, $P = 4$ should comply the following condition:

$$\sum_{\forall i \in \{1,\ldots,P\}} \delta_{if} = 1 \qquad P = 4,\ \forall f \in FN. \tag{20}$$

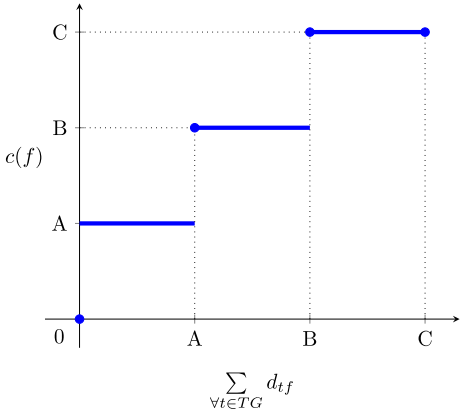


Fig. 1. Fog node capacity function.

Finally, the capacity of each FN is defined as:

$$c(f) = \delta_{1f} \cdot 0 + \delta_{2f} \cdot A + \delta_{3f} \cdot B + \delta_{4f} \cdot C \qquad \forall f \in FN. \tag{21}$$

Overall, restrictions (14) to (21) replace the set of constraints (8), used in the model to determine the capacity value for each FN f, such that it will always be higher than the covered demand (otherwise, another FN is selected to cover the unsatisfied service requirements).

2.2. Linearization of the euclidean norm

In this subsection, the linearization procedure for (9) is shown. The proposal of (Camino et al., 2016) is followed to linearize the computation of the Euclidean distance for continuous points in $\mathbb{R}^2$. The basis is to discretize the directions of the Euclidean plane, which is characterized by the continuous domain $[0,2\pi]$, by n_d directions of size $\frac{2\pi}{n_d}$. Thus, the ith discretized direction is the following unit vector U_i:

$$U_i = \left[\cos\left(\frac{2(i-1)\pi}{n_d}\right), \sin\left(\frac{2(i-1)\pi}{n_d}\right)\right]^T \quad \forall i \in \{1, \ldots, n_d\},$$

being $\|U_i\| = 1$

To verify whether two points $\mathbf{p_A} = (x_A, y_A)$ and $\mathbf{p_B} = (x_B, y_B)$ are closer than a given distance d_{TFmax}, we check that all the projections of the $\mathbf{p_A}$–$\mathbf{p_B}$ vector on these directions are lower than $d_{TFmax} \cdot \cos(\theta_{\max})$, being $\theta_{\max} = \frac{\pi}{n_d}$.

$$\begin{aligned}(x_A - x_B) \cdot \cos\left(\frac{2(i-1)\pi}{n_d}\right) + (y_A - y_B) \cdot \sin\left(\frac{2(i-1)\pi}{n_d}\right) \\ \leq d_{TFmax} \cdot \cos(\theta_{\max}) \quad \forall i \in n_d, \quad \forall t \in TG, \quad \forall f \in FN.\end{aligned} \tag{22}$$

Moreover, we have to linearize the following proposition:

$$\text{If } u_{tf} = 1 \Rightarrow \text{distance(t, f)} \leq d_{TFmax} \text{ is TRUE,} \tag{23}$$

which is equivalent to:

$$\text{distance}(t,f) - \text{MaxD} \cdot (1 - u_{tf}) \leq u_{tf} \cdot d_{TFmax}, \tag{24}$$

being MaxD the maximum distance between two locations.

Thus, from inequalities 22 and 24, the following constraint is obtained:

$$\begin{aligned}(x_A - x_B) \cdot \cos\left(\frac{2(i-1)\pi}{n_d}\right) + (y_A - y_B) \cdot \sin\left(\frac{2(i-1)\pi}{n_d}\right) \\ - \text{MaxD} \cdot (1 - u_{tf}) \quad \leq u_{tf} \cdot d_{TFmax} \cdot \cos(\theta_{\max}).\end{aligned} \tag{25}$$

2.3. Linearization of the TG-FN assignments

The constraint set in (10) relates the part of TG t demand served by a FN with the binary variable u_{tf}, which determines if this relation really exists. Thus, (10) involves the following implications:

$$\text{if } u_{tf} = 0 \Rightarrow d_{tf} = 0 \qquad \forall t \in TG, \forall f \in FN \tag{26}$$

$$\text{if } u_{tf} = 1 \Rightarrow d_{tf} > 0 \qquad \forall t \in TG, \forall f \in FN \tag{27}$$

which are equivalent to the following constraints, being ϵ an arbitrary small value:

$$d_{tf} \leq C \cdot u_{tf} \qquad \forall t \in TG, \forall f \in FN \tag{28}$$

$$d_{tf} \geq \epsilon \cdot u_{tf} \qquad \forall t \in TG, \forall f \in FN \tag{29}$$

Repeating the same procedure for 11, the following must be linearized:

$$\text{if } v_f = 0 \Rightarrow \sum_{\forall t \in TG} d_{tf} = 0 \qquad \forall f \in FN \tag{30}$$

$$\text{if } v_f = 1 \Rightarrow \sum_{\forall t \in TG} d_{tf} > 0 \qquad \forall f \in FN \tag{31}$$

consequently equivalent to the restrictions below:

$$\sum_{\forall t \in TG} d_{tf} \leq C \cdot v_t f \qquad \forall f \in FN \tag{32}$$

$$\sum_{\forall t \in TG} d_{tf} \geq \epsilon \cdot v_f \qquad \forall f \in FN \tag{33}$$

Therefore, (10) has to be replaced by restrictions (28) and (29), while (11) have to be replaced by constraints (32) and (33).

3. Placement algorithm: Hybrid Simulated Annealing

Capacitated FLPs are considered part of the NP-hard problem set (Silva and la Figuera, 2007; Qin et al., 2015; Farahani et al., 2012; Wu et al., 2006; Zhu et al., 2010). The FNPP tackled in this paper, being a combination of two FLPs problem types could be derived to be NP-hard. In a nutshell, it implies the analysis of all the possible FN-TG combinations in order to find the minimum cost solution. What is more, given the latency constraints and the need to satisfy all TG demands in a capacity-dependent cost model, the combinations cannot be split to reduce computation time.

On the other hand, there is still an absence of working knowledge and operational data regarding 5G user behavior, future traffic patterns and service trends, in a FC, NFV, 5G ecosystem. Therefore, predicting the number of FNs for a given service area is a nearly impossible task. What is certain, is that ultra-dense networking and 5G stringent requirements will push the amount of FNs to thousands in just a city. Although the MILP formulation makes the FNPP solvable by any available solver (e.g. GLPK), for those scalability requirements the problem difficulty increases abruptly and so does the execution times and required computational resources. Taking this into consideration, a heuristic method based on the simulated annealing (SA) algorithm was developed for the placement strategy: **Hybrid Simulated Annealing (Hybrid-SA)**.

SA has been already used to solve FLPs (Qin et al., 2012, 2015; Ho and Wu, 2012). Overall, selecting SA as a solution was a decision based on its flexibility to solve combinatorial problems when compared to other solutions such as the Lagrangian method and branch and bound algorithms. In addition, SA has been already tested and compared to other heuristics when solving FLPs, showing excellent results in both performance and solution quality when compared to best known or heuristic-generated values (Qin et al., 2015; Ho and Wu, 2012; Arostegui et al., 2006; Goiri et al., 2011; Delmaire et al.,).

In spite of its benefits, SA showed a non-convergent behavior during our experiments. The obtained solutions were widely diverse in terms of cost and number of FNs despite varying the cooling parameters and iteration counters. To solve this problem and improve the obtained results, we decide to develop an SA-based strategy mixing some of the core ideas behind efficient methods such as Tabu Search (TS). The idea was to inherit the flexibility of SA and combine it with the use of memory structures as done in TS (Glover and Kochenberger, 2003), and local search techniques. The indicated method allowed us to improve our experimental results when compared to the traditional SA implementation (see Section 4). Further details about the Hybrid-SA method developed are described in Section 3.1.

3.1. Heuristic description

With the purpose of reducing computation time without loss of generality and accuracy, the algorithm starts by finding the isolated TGs. Isolation occurs when a TG has no other TG closer than $D_{\max}$, which means that it should necessarily be upgraded to FN.

This concept was extended to Pre-Optimized TG Areas (PTAs) which resulted in a significant reduction of the solution search space. A PTA is defined as any TG group where regardless of the TG upgraded to FN there is no impact on the solution quality. The reason is that every TG is within the coverage area of any other TG while remaining isolated to other TGs outside the PTA. Since our deployment strategy was

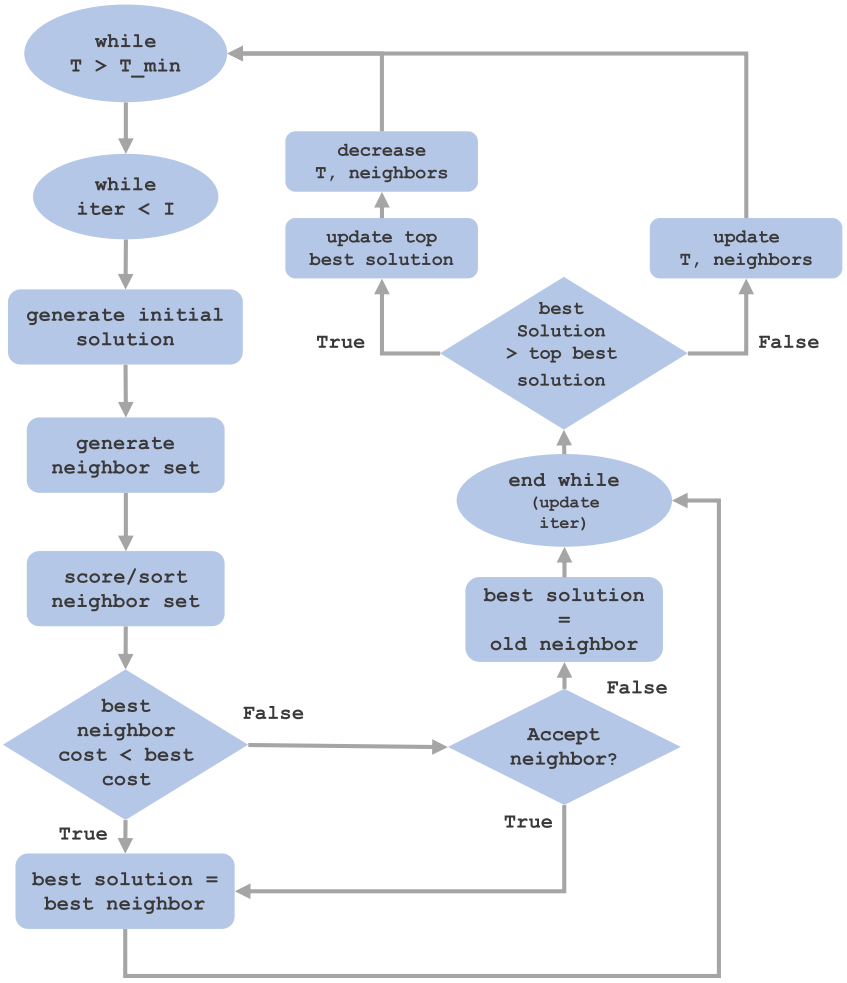


Fig. 2. Hybrid Simulated Annealing flow diagram.

thought for the FNs to be placed in both rural and urban areas where the TG density is expected to be lower, dealing with isolated TGs and PTAs in advance improves the overall performance of the proposed heuristic.

The pseudo-code of the proposed algorithm is showed in Fig. 2. The first critical step is to create a good enough initial solution. For this purpose we develop a greedy strategy where random TGs are upgraded to FNs, taking into account capacity and latency limitations and ensuring that no FN is assigned unneeded capacity. Secondly, a set of neighbor solutions (called individuals) based on this initial step is obtained. The neighbor set contains a predefined number of individuals and is divided in a subset of solutions based on good, bad and randomly generated solutions. The overall idea was to widely explore the search space in each iteration.

Generating new solutions based on good previous individuals ensures the convergence of the algorithm into the best placement locations found (in terms of overall cost and number of FNs). For this purpose, it is crucial to ensure that a new individual resembles the previous obtained one. This is performed by selecting the TGs to be upgraded to FNs within the vicinity of the old selected FNs. Additionally, as a diversification strategy (as in Tabu Search techniques), random and bad solutions are generated to visit unexplored areas of the solution space. As the system "cools down", the number of neighbors generated in each iteration changes as part of the intensification process. As a result, less bad and random solutions are created while the number of good solutions is increased as long as there are cost improvements. If after an iteration cycle for a given temperature, the cost is better than the best cost ever recorded (short-term memory structure part of the intensification process), a penalty function based on a random probability decreases the number of neighbors, and the speed of temperature reduction. The function probability gets higher as the temperature declines. Consequently, the system "cools down" quickly when there are continuous cost improvements and convergence to the global optima, while, otherwise, it slowly changes the temperature and aggressively finds more solutions.

To evaluate each solution, a scoring method was developed. Both the cost and the number of FNs had to be taken into account, but their values were in different orders of magnitudes. The solution was to normalize the values using logarithms and then estimate the distance from both values, as a coordinate pair, to the coordinate origin (0, 0). The obtained value was then use to evaluate the solutions found in each iteration and score them accordingly.

To reduce computation times, facility locations are assumed to be co-locations of existing TGs. This approach offers a near-optimal solution in acceptable running times without extreme usage of computing resources for a fairly large number of TGs. Such assumption is supported by two main facts: capital and operational investments could be minimized by reusing already existing infrastructure and site conditions (space, networking and powering lines, etc.) on the high service demand locations. Additionally, placing the infrastructure the closest possible to the aggregation points on the service access layer, will significantly decrease end-to-end latency.

The set of input parameters required is showed in Table 1. From Eq. (1), D_{max} calculation is of main relevance. As the aim is to

Table 1
Input parameters for the placement algorithm.

Parameter	Meaning
D_{max}	Maximum allowed distance between a TG and its serving FN
TG	Set of TGs coordinates
FNcapacities	Fog nodes capacities specified as small-sized, medium-sized and large-sized FNs
MapGrid	Territory where TGs are located

reduce latency from the service infrastructure (FNs) to the cell traffic aggregation points (TGs in this context), the delay was arbitrarily selected to be 1 ms, 3 ms and 5 ms. These delay values comply with the 5G latency requirements for a wide variety of use case scenarios.

Fig. 3 illustrates a simulated territory of interest (10,000 km^2). TGs are distributed in three *cities* and randomly in rural areas, consequently emulating a reasonably realistic distribution of demand points, where urban areas present higher traffic density.

3.2. Complexity analysis

Since the core of our heuristic is the SA method, the traditional implementation goes through t temperature steps where the related complexity is $O(\log t)$. For each t the search is executed a fixed number of iterations and generates $O(n)$ neighbor solutions. The solution generation method populates the neighbor set. For this function the worst case are the "solution-based individuals", as they loop through previous generated solutions, FN by FN ($O(f)$, being f the number of FNs in the baseline solution), in search for random candidates (TGs suitable to be upgraded to FNs) within each FN coverage area. This iterative process is directly linked to the maximum number of TGs, conventionally called M, to be found under the most populated coverage area. M is determined by running a greedy algorithm (see Section 4) while assigning to each FN the maximum available capacity. It is easy to conclude that M cannot be found beforehand and that the overall algorithm complexity should be formulated based on it. Based on this analysis the complexity can be specified as $O(n \cdot f \cdot M \cdot \log t)$. The initial value of the number of neighbor solutions is relatively small and it is reduced as the system converges. Therefore, the overall algorithm complexity can be defined as $O(f \cdot M \cdot \log t)$.

Table 2
Input parameters values.

D_{max} (*km*)	TG number	Cap. L-FN	Cap. M-FN	Cap. S-FN
3	100	40	30	21
	200	59	41	32
	300	80	51	40
	400	105	74	58
	500	150	101	78
9	100	74	51	38
	200	138	94	71
	300	209	146	102
	400	291	205	153
	500	348	238	179
15	100	75	52	41
	200	154	101	79
	300	240	157	119
	400	326	214	160
	500	392	268	193

4. Case study and results

In order to compare the performance of the placement strategy proposed, a traditional SA implementation, the Hybrid-SA approach and the MILP were run for three latency values: 1 ms, 3 ms and 5 ms. For the case study of mobile Radio Access Networks (RANs) and a Cloud-RAN (C-RAN) architecture, virtualized Baseband Units (BBUs) (Abdelwahab et al., 2016), are to be placed at the FNs. Consequently, from (Musumeci et al.,; Chang et al., 2016) a backhaul transmission delay for LTE networks is known to be around 250 μs Therefore, for 5G networks and the proposed latency values, D_{max} was estimated to be 3 km, 9 km, 15 km (for transmission times of 31 μs, 93 μs, 156 μs). The input list can be observed in Table 2.

A map grid of 100 km × 100 km was used with a set of TG ranging from 100 to 500 TGs (with a 100 TGs increase step in each simulation).

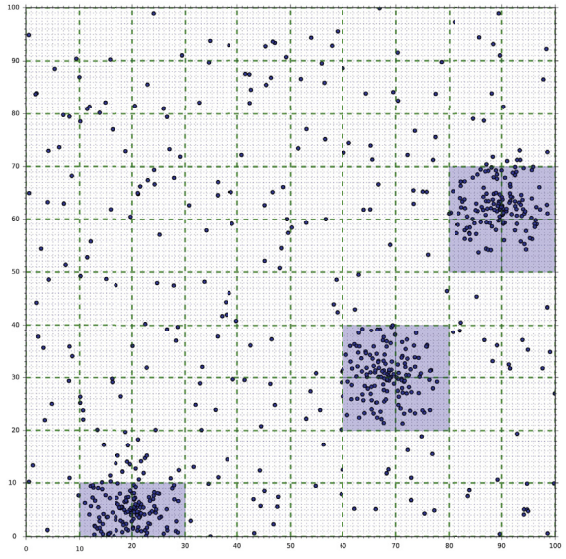


Fig. 3. Traffic generators randomly distributed in three *cities*.

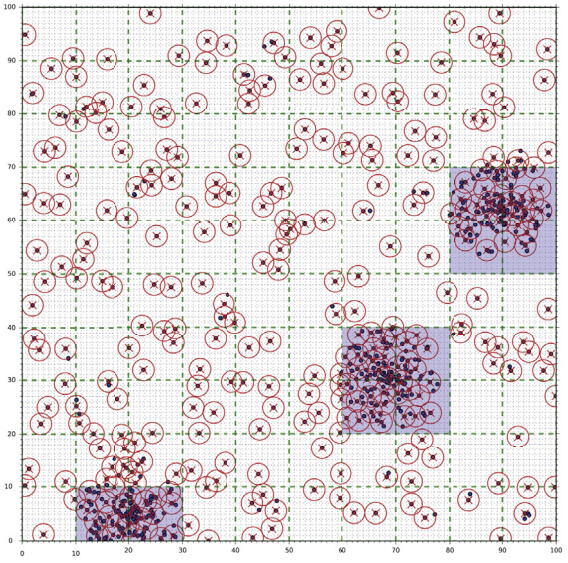


Fig. 4. Solution obtained after running the algorithm.

For the heuristic, the *temperature* ranged from 1.0 to 0.001 with a step size for the fast temperature reduction of $\alpha = 0.8$ and $\alpha = 0.9$ for the lower stepping process. The number of iterations per temperature was set to 10 for the Hybrid-SA as several solutions are created and evaluated in each iteration (the number of neighbor individuals in each iteration was set to 8). The iteration counter for the traditional SA was set to 100. This value was empirically determined, aiming to ensure a fairly similar number of iterations when compared to the Hybrid-SA proposed (in fact 100 is a bit higher to compensate SA lack of accuracy). All simulations were run in a computer with a 2.60 *Hz* 8-core CPU (*x*64 architecture) and 32 GB RAM. Pyomo (Hartand et al., 2011, 2017) was the python-based package selected to solve the optimization model proposed in Section 2, along with GLPK as underlying solver.

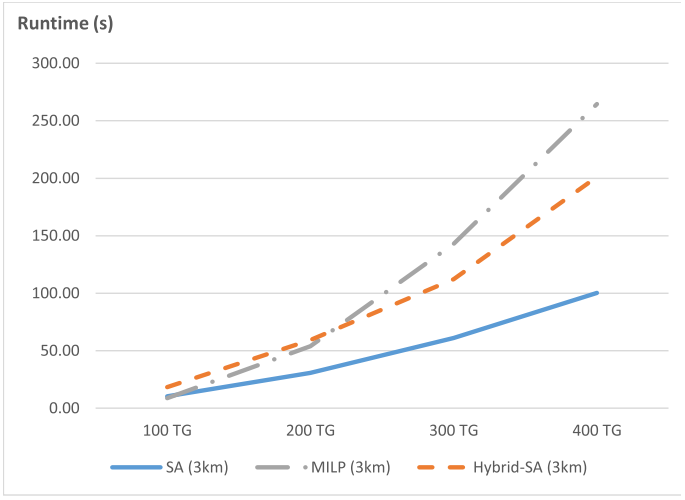
To obtain the FN capacities showed in Table 2, an additional greedy algorithm was developed. It iteratively upgrades to FN the TG with the most populated coverage area (given $D_{\max}$), and keeps on until no TG remains uncovered. As a result, the allowed FN capacities are found for any particular solution. Such greedy algorithm was run several times for each simulation setting described above. Consequently, the final capacity values for the heuristics and the mathematical model were obtained through the statistical analysis of the results.

Fig. 4 displays a final solution after running the heuristic. It can be observed how every TG is covered (FNs are depicted as × and the surrounding circles are the coverage area of $D_{\max}$ radius).

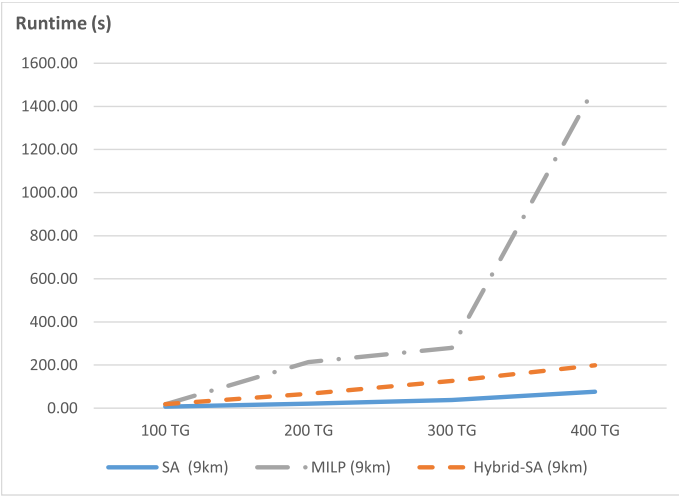
To validate the results, both the Hybrid-SA and the SA were run ten times for every $D_{\max}$ and TG combination. Meanwhile, for each $D_{\max}$ value, the number of TGs was increased as mentioned above, aiming to calculate the execution time and the number of FNs of the optimal solution found. The findings are presented in Figs. 5 and 6.

It can be noticed in Fig. 5 quite a difference in the running times of both heuristics and the MILP model. Despite the steady surge in the first stages for all cases (while increasing the number of TGs), the mathematical model has a nearly impossible task in obtaining the optimal solution for $D_{\max} = \{9, 15\}$ km and TG = 500 (see Fig. 5). Therefore, the experimental results are just shown from 100 TGs to 400 TGs in both Figs. 5 and 6. In fact, the heuristics are able to found a near-optimal solution in significantly less time and with a maximum of a few FNs gap as shown in Fig. 6.

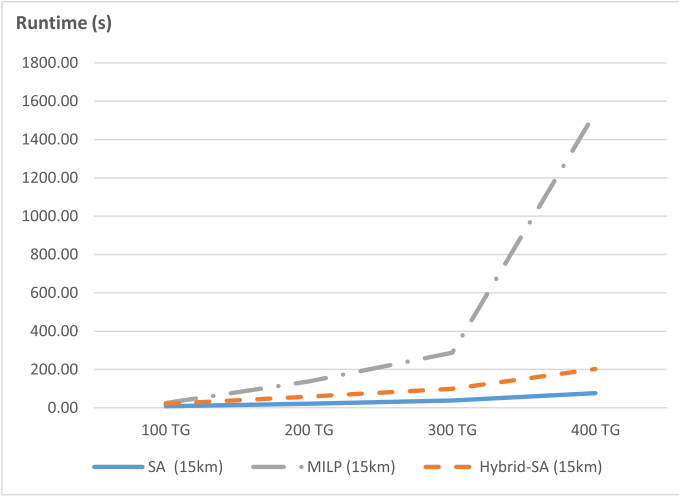
The MILP model execution time rapidly steeps to huge values after reaching 400 TGs, due to the exponential growth in the number of feasible solutions. In contrast, both heuristics running delay climbs regularly



(a) Execution times for $D_{max} = 3\ km$

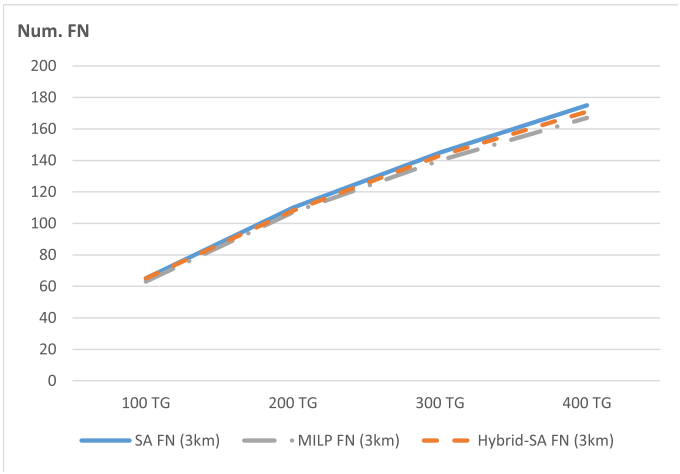


(b) Execution times for $D_{max} = 9\ km$

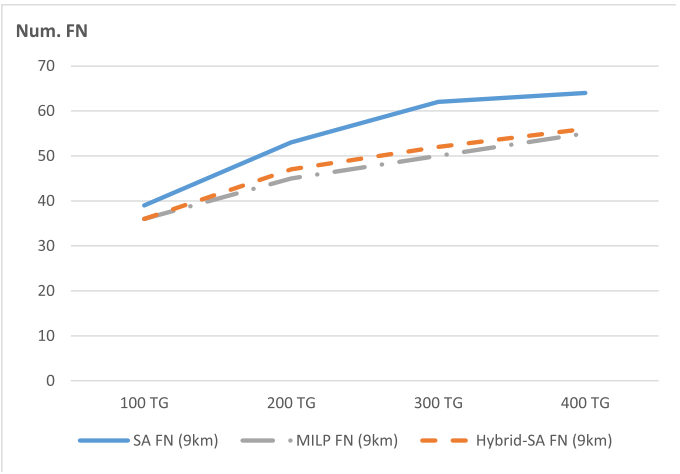


(c) Execution times for $D_{max} = 15\ km$

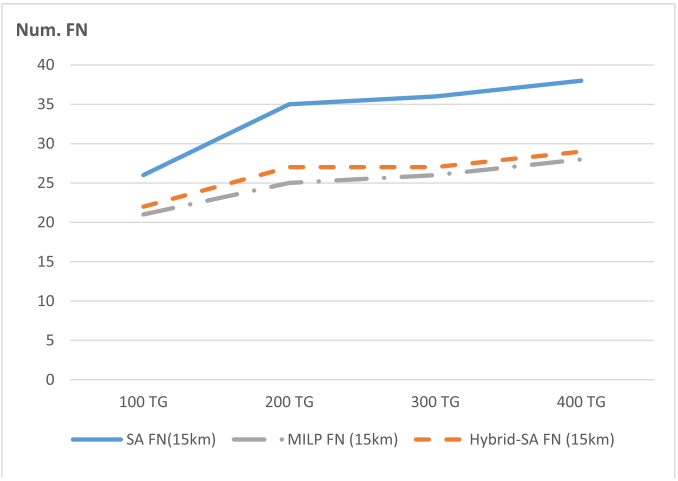
Fig. 5. Execution times for the SA, Hybrid-SA and MILP.



(a) Number of FNs deployed for $D_{max} = 3\ km$



(b) Number of FNs deployed for $D_{max} = 9\ km$



(c) Number of FNs deployed for $D_{max} = 15\ km$

Fig. 6. Number of FNs deployed by the SA, Hybrid-SA and the MILP model.

throughout the TGs experimentation set. Due to the latency constraint variation, the number of FN decreases as $D_{\max}$ rises. The reason is that the FN coverage area becomes larger, thus less FNs are required to cover existing TGs.

Regarding the performance of both heuristics compared to the exact model when minimizing the number of FNs, the Hybrid-SA shows clear improvements at the cost of an increase in the execution time. Since the goal of our study is to place physical infrastructure, the placement strategy is to be run during the planning phase of the deployment and thus this is not considered an issue. However, we strongly believe that the Hybrid-SA performance could be further improved. A recommendation on this matter is to add a distance parameter when searching for the FN candidates in the solution generation method. Since currently it searches for a random TG to upgrade to FN in the vicinity of the old FN, such vicinity could be restricted by a distance parameter iteratively reduced depending on the system temperature. This way, with each search cycle the solution moves faster towards the best TG candidates. This will also impact the number of neighbors generated as the overall systems will increment the frequency of finding better solutions.

The Hybrid-SA performance regarding the number of FN deployed is quite promising. From Fig. 6, the difference between the number of FNs placed by the MILP and the Hybrid-SA approach never surpassed a threshold of even less than 5 FNs. Based on this result, we consider that a thorough analysis of next generation network data about traffic patterns and services, can reduce the gap and improve the efficiency of the Hybrid-SA.

5. Conclusions

In future 5G networks merging Fog Computing and other enabling technologies, the devised number of FNs considering Internet of Things (IoT) scenarios, for instance, will probably scale up to thousands of nodes. This forces strict mathematical formulations as the MILP model aforementioned, to be dropped as possible placement mechanisms. As a result, more flexible methods, such as the proposed heuristic, become the right approach to solve the problem.

It is worth noticing that working knowledge of 5G service infrastructure sizing will ensure optimal results when applying the proposed heuristic. Additionally, since FN capacities will not be arbitrarily calculated but conditioned by available solutions in the market, the obtained results will consequently become of quite practical interest. An added value of the heuristic, is that it could be used initially to find the required capacities for a real distributed demand scenario, thus helping with the infrastructure sizing problem.

Finally, in spite of the promising results presented for our heuristic when compared to the traditional SA implementation and the MILP model, further research is still necessary. Future work should carefully target a more comprehensive set of placement parameters such as location dependent cost and additional candidate sites such as Central Offices and Internet Service Provider infrastructures. Furthermore, given the complexity of the envisioned 5G ecosystem the interdependence with other technologies should be analyzed. Certainly, a thorough analysis of NFV, SDN and Edge Computing implementations will lead to a turning point in the problem formulation, presumably towards a multi-objective and multi-criteria optimization problem.

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 AEI/FEDER, UE.

References

Abdelwahab, S., Hamdaoui, B., Guizani, M., Znati, T., 2016. Network function virtualization in 5g. IEEE Commun. Mag. 54 (4), 84–91.
Arabani, A.B., Farahani, R.Z., 2012. Facility location dynamics: an overview of classifications and applications. Comput. Ind. Eng. 62 (1), 408–420.
Arostegui, M.A., Kadipasaoglu, S.N., Khumawala, B.M., 2006. An empirical comparison of tabu search, simulated annealing, and genetic algorithms for facilities location problems. Int. J. Prod. Econ. 103 (2), 742–754.
E. Balevi, R. D. Gitlin, Optimizing the number of fog nodes for cloud-fog-thing networks, arXiv preprint arXiv:1801.00831.
Barbati, M., 2013. Models and Algorithms for Facility Location Problems with Equity Considerations Ph.D. thesis. Universita degli Studi di Napoli Federico II.
Basta, A., Blenk, A., Hoffmann, K., Morper, H.J., Hoffmann, M., Kellerer, W., 2017. Towards a cost optimal design for a 5g mobile core network based on SDN and NFV. IEEE Trans. Netw. Serv. Manag. 14 (4), 1061–1075.
Blanco, B., et al., 2017. Technology pillars in the architecture of future 5g mobile networks: NFV, MEC and SDN. Comput. Stand. Interfac. 54, 216–228.
Brogi, A., Forti, S., 2017. QoS-aware deployment of IoT applications through the fog. IEEE Internet Things J. 4 (5), 1185–1192.
Byers, C.C., 2017. Architectural imperatives for fog computing: use cases, requirements, and architectural techniques for fog-enabled IoT networks. IEEE Commun. Mag. 55 (8), 14–20.
Camino, J., Artigues, C., Houssin, L., Mourgues, S., 2016. Linearization of euclidean norm dependent inequalities applied to multibeam satellites design. LAAS Publ. 1 (16116), 1–18.
Carlsson, J.G., Carlsson, E., Devulapalli, R., 2016. Shadow prices in territory division. Network. Spatial Econ. 16 (3), 893–931.
Chang, C., Nikaein, N., Spyropoulos, T., 2016. Impact of packetization and scheduling on c-ran fronthaul performance. In: Global Communications Conference (GLOBECOM), 2016 IEEE. IEEE, pp. 1–7.
H. Delmaire, J. A. Daz, E. Fernndez, M. Ortega, Comparing new heuristics for the pure integer capacitated plant location problem.
Dolui, K., Datta, S.K., 2017. Comparison of edge computing implementations: fog computing, cloudlet and mobile edge computing. In: 2017 Global Internet of Things Summit (GIoTS), pp. 1–6.
Elayoubi, S.E., et al., 2016. 5g service requirements and operational use cases: analysis and METIS ii vision. In: 2016 European Conference on Networks and Communications (EuCNC), pp. 158–162.
Farahani, R.Z., SteadieSeifi, M., Asgari, N., 2010. Multiple criteria facility location problems: a survey. Appl. Math. Model. 34 (7), 1689–1709.
Farahani, R.Z., Asgari, N., Heidari, N., Hosseininia, M., Goh, M., 2012. Covering problems in facility location: a review. Comput. Ind. Eng. 62 (1), 368–407.
Glover, F., Kochenberger, G.A., 2003. Handbook of Metaheuristics. Springer US, Boston, MA.
Goiri, I., Kien, L., Guitart, J.J., Torres, J., Bianchini, R., 2011. Intelligent placement of datacenters for internet services. In: 2011 31st International Conference on Distributed Computing Systems ICDCS. IEEE, pp. 131–142.
Gravalos, I., Makris, P., Christodoulopoulos, K., Varvarigos, E.A., 2016. Efficient gateways placement for internet of things with QoS constraints. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–6.
Hartand, W.E., Watson, J., Woodruff, L., 2011. Pyomo: modeling and solving mathematical programs in python. Math. Program. Comput. 3 (3), 219–260.
Hartand, W.E., Laird, C.D., Watson, J., Woodruff, D.L., Hackebei, G.A., Nicholson, B.L., Siirola, D., 2017. second ed. Pyomo–optimization Modeling in python, vol. 67. Springer Science & Business Media.
Ho, Z., Wu, C., 2012. Application of simulated annealing algorithm to optimization deployment of mobile wireless base stations. In: Advances in Computer Science and Information Engineering. Springer, pp. 665–670.
Intharawijitr, K., Iida, K., Koga, H., 2016. Analysis of fog model considering computing and communication latency in 5g cellular networks. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops). IEEE, pp. 1–4.
Krivitski, D., Schuster, A., Wolff, R., 2005. A local facility location algorithm for sensor networks. In: First IEEE International Conference (DCOSS 2005). Springer, pp. 368–375.
F. Musumeci, C. Bellanzon, N. Carapellese, M. T. A. Pattavina, S. Gosselin, Optimal bbu placement for 5g c-ran deployment over wdm aggregation networks, Journal of Lightwave Technology 34(8).
Qin, J., Ni, L., Shi, F., 2012. Combined simulated annealing algorithm for the discrete facility location problem. Sci. World J. 576392, 7 pages. https://doi.org/10.1100/2012/576392.
Qin, J., Xiang, H., Ye, Y., Ni, L., 2015. A simulated annealing methodology to multiproduct capacitated facility location with stochastic demand. Sci. World J. 2015, 1–9.
Rodriguez, J. (Ed.), 2015. Fundamentals of 5G Mobile Networks. Wiley.
Silva, F.J.F., la Figuera, D.S.D., 2007. A capacitated facility location problem with constrained backlogging probabilities. Int. J. Prod. Res. 45 (21), 5117–5134.
Ulukan, Z., Demircioglu, E., 2015. A survey of discrete facility location problems. Int. J. Soc. Behav. Edu. Econ. Bus. Ind. Eng. 9 (7), 2487–2492.
Wang, S., Ran, C., 2016. Rethinking cellular network planning and optimization. IEEE Wirel. Commun. 23 (2), 118–125.
Wang, S., Zhao, W., Wang, C., 2015. Budgeted cell planning for cellular networks with small cells. IEEE Trans. Veh. Technol. 64 (10), 4797–4806.
Wu, L., Zhang, X., Zhang, J., 2006. Capacitated facility location problem with general setup cost. Comput. Oper. Res. 33 (5), 1226–1241.
Xiang, W., Zheng, K., Shen, X., 2017. 5G Mobile Communications. Springer International Publishing.
Yannuzzi, M., et al., 2017. A new era for cities with fog computing. IEEE Internet Comput. 21 (2), 54–67.
Zhang, W., Lin, B., Yin, Q., Zhao, T., 2017. Infrastructure deployment and optimization of fog network based on MicroDC and LRPON integration. Peer-to-Peer Netw. Appl. 10 (3), 579–591.
Zhou, S., Lee, D., Leng, B., Zhou, X., Zhang, H., Niu, Z., 2015. On the spatial distribution of base stations and its relation to the traffic density in cellular networks. IEEE Access 3, 998–1010.
Zhu, Z., Chu, F., Sun, L., 2010. The capacitated plant location problem with customers and suppliers matching. Transport. Res. E Logist. Transport. Rev. 46 (3), 469–480.

Alejandro Santoyo-Gonźalez: Received his BSc. degree in Telecommunication and Electronic Engineering from the Havana University of Technology Jose Antonio Echeverria (Havana, Cuba). He is currently a PhD student and researcher at the Universitat Politècnica de Catalunya (UPC). Experienced as datacenter engineer and solution manager of a leading multinational telecommunications company, his main research interests include NFV, SDN, Edge Computing and 5G networking.

Cristina Cervelló-Pastor: Received her MSc and Ph.D degree in Telecommunication Engineering, both from the Barcelona School of Telecommunications Engineering (ETSETB), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. She is an Associate Professor and the Head of the Dept. of Network Engineering of UPC. Being part of BAMPLA research group she has been responsible and actively participated in diverse national and European competitive projects (NOVI, FEDERICA, ATDMA, A@DAN, Euro-NGI, Euro-FGI, EURONF) and private funding R&D projects. In parallel she has published diverse papers in national and international journals and conferences and she has been supervising thesis in the field of management, optimal resource allocation, topology discovery and routing in SDN/NFV and 5G.